EDV in Medizin und Biologie **16** (4), 117–120, ISSN 0300-8282

Alignment and simple stratification in clinical studies

Michael Schemper

Summary

Stratification is frequently used instead of parametric modelling in analyses of clinical trials to reduce bias and to increase power. Avoiding the adverse effect of extensive stratification – the power decreases with an increasing number of strata – LEHMANN (1975) proposed an alignment technique with impressive asymptotic efficiencies. As the usefulness of alignment with analyses of clinical trials has never been investigated, the author tried to reach his conclusions by means of an extensive Monte Carlo study, contrasting the unstratified and stratified Mann-Whitney test and the alignment test by their size and power in many situations likely to occur in practice. Similarly, the effect of increasing censoring on GEHAN (1965)-type generalizations of the three tests was also studied.

Samples were taken from normal and exponential distributions with and without censoring and for equal and unequal strata sizes. The main conclusion is that alignment tests can improve the power of stratified analyses of clinical studies considerably if average strata sizes are below or equal to 8 and censoring below or equal to 16 %. The size of the three tests was near the nominal level $\alpha = 0.05$ and $\alpha = 0.01$.

Zusammenfassung

Anstelle parametrischer Modellierung in der statistischen Analyse klinischer Studien wird häufig Stratifizierung verwendet, mit dem Ziel größerer Teststärke und geringerer Verfälschtheit. Um den ungünstigen Effekt starker Stratifizierung zu umgehen – bei steigenden Strataanzahlen nimmt die Teststärke ab –, schlug LEHMANN (1975) eine Alignment-Technik vor, mit eindrucksvollen asymptotischen Effizienzen. Wie weit sich die Alignment-Technik zur Analyse klinischer Studien eignet, ist bisher nicht untersucht worden. Mit Hilfe einer umfangreichen Monte-Carlo-Studie, in der der nichtstratifizierte und der stratifizierte Mann-Whitney-Test sowie der Alignment-Test einander hinsichtlich eingehaltenem nominalem Testniveau und der Teststärke gegenübergestellt wurden, konnten Schlüsse für typische Anwendungssituationen gezogen werden. Weiters wurde der Einfluß steigender Zensierung auf Verallgemeinerungen der drei Verfahren für zensierte Daten (entsprechend GEHAN,1965) untersucht. Zufallsstichproben wurden aus Normal- und Exponentialverteilungen mit und ohne Zensierung sowie für gleiche und ungleiche Strataumfänge gezogen. Als wesentlicher Schluß kann gelten, daß Alignment-Tests die Teststärke für stratifizierte Analysen in klinischen Studien wesentlich verbessern können – bei durchschnittlichen Strataumfängen bis 8 und bis 16 % zensierten Beobachtungen. Das nominelle Signifikanzniveau $\alpha = 0{,}05$ und $\alpha = 0{,}01$ wurde von den 3 Tests immer eingehalten.

1. Introduction

Rank methods have a distinct advantage over classical normal theory procedures in comparative experiments with two or more treatments due to their insensitivity to gross errors and extreme observations. Furthermore, they do not require a metric scale for the variate. The asymptotic relative efficiencies (A.R.E.) of the familiar Mann-Whitney, »paired« Wilcoxon or Kruskal-Wallis tests hold up quite well under the assumption of a normal distribution – $3\ \pi$ relative to corresponding t and F tests. However, in many statistical comparisons the experimental subjects must be divided into homogenous strata to correct for imbalances in the distribution of covariates and to obtain increased precision and power.

Although stratification – to avoid any doubtful modelling assumptions – can be embedded in either parametric or nonparametric analyses, only the latter are considered here.

HODGES and LEHMANN (1962) pointed out that the rather low A.R.Es of separate ranking (within strata)-procedures (e.g. the sign test with A.R.E. of $2/\pi$ or Friedman's test) are due to the absence of intrablock comparisons. Their alignment procedures make it possible to carry out stratified testing with A.R.Es similar to those given above for the Mann-Whitney test under normality. TARDIF (1981) attested satisfactory asymptotic behavior to 'ranking after alignment' – procedures and MEHRA and SARANGI (1967) and SARANGI and MEHRA (1969) generalized the procedure to compare k treatments within blocks allowing for arbitrary and unbalanced frequencies within the resulting cells. LEHMANN (1975) again recommends alignment procedures for stratified analyses, quoting A.R.Es derived in the two papers by MEHRA and SARANGI.

The performance of alignment procedures in the small and medium samples encountered in practice has not previously been examined. The author therefore undertook extensive simulation studies comparing two treatment groups in the presence of block or stratum effects under various conditions.

2. Methods

2.1 Description of tests

Suppose two treatments ($j = 1, 2$) are compared in an experimental design with n strata ($i = 1, \ldots, n$) and m_{ij} (≥ 1) observations in the (i, j)-th cell. The total number of observations is N ($= \sum_{i=1}^{n}\sum_{j=1}^{2} m_{ij}$), the treatment totals are $N_j = \sum_{i=1}^{n} m_{ij}$ and the size of the i-th stratum is $m_i = m_{i1} + m_{i2}$.

The m_{ij} random variables X_{ijl} ($l = 1, 2, \ldots, m_{ij}$) are independently and identically distributed with a common cumulative distribution function $F_{ij}(X) = F_j(X + \xi_i)$ where ξ_i are unknown stratum effects. The hypothesis of no difference between the treatment effects can be expressed as $H_0: F_1 = F_2$.

Comparing two arbitrary observations of the total sample, X_g and X_h, we define

$$U_{gh} = \begin{cases} 1 \\ -1 \\ 0 \end{cases} \text{ if } (X_g - X_h) \gtrless 0 \text{ and Min } (X_g, X_h) \text{ uncensored} \quad (1)$$
$$\text{otherwise}$$

The tests included in this study are briefly described below.

Unstratified comparison:

The statistic of the test by MANN and WHITNEY (1947) and of its generalization for censored data by GEHAN (1965) is given by

$$W = \sum_{g<h} U_{gh}$$

where indices g and h refer to all observations of treatment 1 and treatment 2 respectively. The sum therefore contains N_1N_2 comparisons. The expectation E(W) is zero and the corresponding variance

$$\text{Var}(W) = N_1N_2 \sum_{f=1}^{N} \sum_{g=1}^{N} \sum_{h=1}^{N} U_{fg}U_{fh}/(N(N-1))$$

reduces to $\text{Var}(W) = N_1N_2(N+1)/3$ in the case of no ties and no censoring.

Stratified comparison:

The statistic for the simple stratified case is given by

$$W^s = \sum_{i=1}^{n} \sum_{g<h} U_{gh}$$

where indices g and h refer to all observations of treatment 1 and treatment 2, respectively, within the i-th stratum. The total sum therefore contains $\sum_{i=1}^{n} m_{i1}m_{i2}$ comparisons. The expectation $E(W^s)$ is zero and the corresponding variance

$$\text{Var}(W^s) = \sum_{i=1}^{n} m_{i1}m_{i2} \sum_{f=1}^{m_i} \sum_{g=1}^{m_i} \sum_{h=1}^{m_i} U_{fg}U_{fh}/(m_i(m_i-1))$$

reduces to $\text{Var}(W^s) = \sum_{i=1}^{n} m_{i1}m_{i2}(m_i+1)/3$ in the case of no ties and no censoring.

Comparison after alignment:

»Alignment« essentially means the removal of the strata effects ξ_i from the observations by subtracting a reasonable function of the observations of a stratum – usually the mean or the median – from each observation in a stratum, prior to statistical comparison of the treatments. The (censored) median chosen here is better suited to nonparametric analysis.

»Scoring after alignment« was carried out instead of »ranking after alignment«, which simplifies the treatment of censored observations. The score S_g of the g-th observation of the total sample is therefore defined by

$$S_g = \sum_{h=1}^{N} U'_{gh}$$

where U'_{gh} is U_{gh} of (1), but calculated from the aligned observations. The statistic $W^a = \sum_{i=1}^{n} \sum_{l=1}^{m_{ij}} S_{i1l}$, its expectation

$$E(W^a) = \sum_{i=1}^{n} m_{i1}/m_i \sum_{j=1}^{2} \sum_{l=1}^{m_{ij}} S_{ijl} \text{ and its variance}$$

$$\text{Var}(W^a) = \sum_{i=1}^{n} m_{i1}m_{i2}/(m_i-1)) \left[\sum_{j=1}^{2} \sum_{l=1}^{m_{ij}} S_{ijl}^2/m_i - \left(\sum_{j=1}^{2} \sum_{l=1}^{m_{ij}} S_{ijl}/m_i\right)^2\right]$$

are defined similarly to LEHMANN (1975), S_{ijl} denoting the score alloted to the l-th observation within stratum i and treatment j.

Under H_0 the statistics W, W^s and W^a asymptotically follow normal distributions with expectations and variances as given.

2.2. *Monte Carlo sampling procedure*

In order to compare the performance of the three tests, data were generated from standard normal distributions and exponential distributions with and without censoring. Any size or power estimate is based on 1000 simulated clinical trials.

When sampling from normal distributions with unit variance, the means were taken equal to study the size. In order to study power, a »small« treatment effect (TES) meant a difference in means of 0.3, a »big« one (TEB) 0.6. Zero (SEZ), »small« (SES) and »big« (SEB) strata effects meant a maximum difference in strata means of 0 (SEZ), 1 (SES) and 2 (SEB). The means of the intermediate strata were equally spaced between the defined minima and maxima. The additivity of treatment and strata effects was assumed.

Similar to LININGER et al. (1979), exponential failure times X were generated with hazard $\exp(\theta_j + \delta_i)$, where $\theta_1 = 0$ and $\theta_2 = 0$ (no treatment effect), $\theta_2 = \ln 1.5$ (TES), $\theta_2 = \ln 2$ (TEB). The effects of the n strata where either defined as $\delta_i = 0$ (SEZ), as $0, \frac{1}{n-1}(\log 2, 2\log 2, \ldots, (n-1)\log 2)$ for SES, or as $0, \frac{1}{n-1}(\log 8, 2\log 8, \ldots, (n-1)\log 8)$ for SEB. Loglinear equally spaced effects were thus assumed for the n strata, as was the multiplicativity of treatment and strata effects on the hazard.

In the experiments with exponential failure times under censoring, a single, independent uniform $(0, \tau_P)$ censoring mechanism was assumed to operate on all strata and treatment groups. This assumption is appropriate when patients enter a study at random over the interval $(0, \tau_P)$, at the end of which a test is performed on the data. For a given overall proportion censored P (cf. LININGER et al., 1979)

$$P = \frac{1}{\tau_P} \sum_{i=1}^{n} \sum_{j=1}^{2} \frac{m_{ij}}{N} \lambda_{ij} (1 - \exp(\lambda_{ij}\tau_P))$$

τ_p is solved iteratively, where λ_{ij} is the hazard of treatment j within stratum i. For P = ⅙, ⅓ and ⅔ uniform variates $U_{1/6}$ on $(0, \tau_{1/6})$, $U_{1/3}$ on $(0, \tau_{1/3})$ and $U_{2/3}$ on $(0, \tau_{2/3})$ were generated for each patient as intervals of his observability. Then the times actually observed for each patient are given by min (X,

$U_{1/6}$), min (X, $U_{1/3}$) and by min (X, $U_{2/3}$). Thus in the 4 nonindependent simulations of P = 0, ⅙, ⅓, ⅔, the variability in studying the effect of increased censoring is reduced.

3. Monte Carlo results

The size of the three alternative tests was investigated in almost all the situations in which the power was simulated. However the results are not given in detail as it is general knowledge that the MANN-WHITNEY (1947) and GEHAN (1965) tests, whether stratified or not, adhere to nominal α-levels (of α = 0.05 and α = 0.01) rather accurately. The same conclusion holds true for aligned comparisons, a result that could have been anticipated by TARDIF's (1981) elaborations.

In the specific situation of N = 32, n = 16, the simple stratified test reduces to the sign-test with 16 pairs. The normal approximation to the distribution of the sign-test requires a much larger number of pairs. The size was kept for α = 0.01 for both normally and exponentially distributed data. This was not the case for α = 0.05 : 0.09 was observed in this particular situation and the simple stratified test. The respective power estimates are therefore omitted from Tables 1 and 2.

With this exception, none of the three tests is superior with regard to its validity – at least in the situations N ≥ 32 and P ≤ 0.67.

3.1 Results for balanced samples with complete observations

In comparing the power of the three tests under the alternatives given in 2.2 for samples from normal and exponential distributions and under balanced treatment and strata frequencies – $m_{i1} = m_{i2}$ and $m_i = N/n$ – all possible combinations of N = 32, 64, 128 and 192, n = 4, 8, 16, 32 for strata effects zero (SEZ), »small« (SES) and »big« (SEB) and treatment effects small (TES) and big (TEB) were investigated at α = 0.05 and 0.01. The results for SES were omitted from Tables 1 and 2 as they typically lay between SEZ and SEB. As the relative behavior of the three tests was rather similar at α = 0.05 and α = 0.01 the results for the latter are not given either.

The results indicate that the alignment test gives remarkably improved power over the simple stratified test only for strata sizes of 8 or below, or correspondingly, the power of the simple stratified test declines remarkably only for strata sizes of 8 or below, while the alignment test almost preserves the power for any strata sizes at the level of the unstratified test in corresponding situations of no block effect. This property of the alignment test observed in Table 1 and probably valid more generally for identical distributional shapes of the strata cannot be observed in Table 2. If block effects exist, then the power of the alignment test does not reach the power of the corresponding unstratified test with no strata effects.

Transforming the observations before alignment by exp(–x) to a uniform distribution did raise the power of the alignment test by 1–2 percent. Adjustment for differing variation (cf. LEHMANN 1975, p. 273; mean absolute deviation from median) further meant that alignment with SEB was not less powerful (as shown in Table 2) for strata sizes $m_i \geq 8$ than simple stratification with SEZ. For small strata of $m_i \leq 4$, adjustment by estimated variation within strata even reduced the power considerably. It therefore seems that for small strata – where alignment is still more powerful than simple stratification – little can be done to avoid a moderate loss in power if strata differ in variation or skewness.

It is unnecessary to add, that the unstratified test is principially weaker in the presence of strata effects than either the simple stratified or the alignment test. Differences between

Table 1. Estimates of power, in percent – based on 1000 simulated completely balanced clinical trials – for unstratified/stratified Mann-Whitney test/alignment test with normally distributed complete observations at α = 0.05.

N	n	m_i	SEZ		SEB	
			TES	TEB	TES	TEB
32	4	8	12/12/13	35/34/36	4/12/13	19/34/36
	8	4	12/10/12	34/26/34	5/10/12	21/26/34
	16	2	13/—/12	35/—/34	6/—/12	23/—/34
64	4	16	21/20/21	63/60/62	9/20/21	43/60/62
	8	8	22/18/21	65/58/64	12/18/21	49/58/64
	16	4	20/16/19	63/52/61	12/16/19	49/52/61
	32	2	19/17/20	64/49/63	11/17/20	49/49/63
128	4	32	37/36/37	91/90/91	20/36/37	79/90/91
	8	16	36/34/36	90/89/90	25/34/36	81/89/90
	16	8	37/32/36	92/88/91	26/32/36	84/88/91
	32	4	37/30/35	91/83/90	25/30/35	85/83/90
192	4	48	52/51/51	98/98/98	32/51/51	93/98/98
	8	24	52/50/52	98/97/98	36/50/52	95/97/98
	16	12	55/53/54	98/98/98	42/53/54	96/98/98
	32	6	54/47/52	98/97/98	40/47/52	96/97/98

Table 2. Estimates of power, in percent – based on 1000 simulated completely balanced clinical trials – for unstratified/stratified Mann-Whitney test/alignment test with exponentially distributed complete observation at α = 0.05.

N	n	m_i	SEZ		SEB	
			TES	TEB	TES	TEB
32	4	8	15/16/16	36/32/36	7/16/13	20/32/32
	8	4	15/12/16	36/26/39	8/12/16	23/26/36
	16	2	17/—/17	37/—/38	10/—/15	24/—/35
64	4	16	29/27/29	65/62/65	16/27/26	47/62/58
	8	8	28/24/28	67/59/66	18/24/26	51/59/61
	16	4	29/21/29	66/51/66	20/21/27	51/51/62
	32	2	28/20/30	65/47/67	16/20/28	52/47/62
128	4	32	50/49/50	92/91/92	31/49/44	80/91/88
	8	16	49/48/50	93/91/93	35/48/45	83/91/89
	16	8	52/47/53	93/90/93	36/47/48	86/90/91
	32	4	50/40/52	92/85/93	36/40/48	85/85/91
192	4	48	68/68/70	100/99/99	49/68/63	95/99/98
	8	24	67/65/68	98/98/98	50/65/61	96/98/97
	16	12	68/64/68	98/98/98	54/64/65	97/98/98
	32	6	67/59/67	98/96/98	52/59/70	96/96/97

Table 3. Estimates of power, in percent – based on 1000 simulated clinical trials with unequal strata sizes – for unstratified/stratified Mann-Whitney test/alignment test with normally (n.d.) and exponentially (e.d.) distributed complete observations at α = 0.05.

N	n	$\bar{m}_i$	SEZ		SEB	
			TES	TEB	TES	TEB
n.d.						
64	16	4	21/14/20	65/45/63	12/14/20	54/45/63
128	16	8	39/28/38	93/77/92	30/28/38	89/77/92
e.d.						
64	16	4	26/16/27	64/42/65	18/16/25	51/42/62
128	16	8	51/38/52	93/79/93	43/38/49	89/79/91

Table 4. Estimates of power, in percent – based on 1000 simulated completely balanced clinical trials – for unstratified/stratified Gehan test/corresponding alignment test with exponentially distributed censored observations at $\alpha = 0.05$ for N = 64 and »big« treatment effect (TEB).

	n	m_i	% censoring 0	1/6	1/3	2/3
SEZ	4	16	65/62/65	60/58/55	52/51/43	30/29/15
	16	4	66/51/66	60/49/54	53/45/42	33/25/17
	32	2	65/47/67	59/44/53	51/38/42	32/21/16
SEB	4	16	47/62/58	43/57/46	38/50/36	25/29/16
	16	4	51/51/62	48/49/51	41/42/37	27/26/14
	32	2	52/47/62	47/44/50	41/37/37	27/20/14

the three tests may become negligible with strong treatment effects and/or big sample sizes, when the power is high anyway.

3.2 Results for unequal strata sizes

Strata of equal size are rarely found in practice. The extent to which conclusions derived from simulations of completely balanced trials have to be corrected for situations of unequal strata is indicated by Table 3. The results for N = 64 are based on strata sizes 14, 8, 8, 4, 4, 4, 4, 2, 2, 2, 2, 2, 2, 2, 2, 2 with an average stratum size of $\overline{m_i} = 4$ and n = 16. The results for N = 128 are based on stratum sizes 32, 20, 16, 12, 8, 8, 8, 4, 4, 4, 2, 2, 2, 2, 2, 2 with an average stratum size of $\overline{m_i} = 8$ and n = 16. Strata and treatment effects are the same as for Tables 1 and 2.

Comparing the results of Table 3 with the corresponding results for N = 64 and 128 and n = 16 in Tables 1 and 2, it seems evident that unbalanced strata frequencies can reduce the power of simple stratified tests remarkably but not that of alignment tests. Thus in unbalanced situations alignment tests may be superior even for average stratum sizes of $\overline{m_i} > 8$.

3.3 Results for balanced exponential samples under censoring

The power of the three tests with exponential, possibly censored failure times – as frequently encountered in clinical trials – is depicted in Table 4 for N = 64, n = 4, 16, 32, for expected proportions censored of 0, ⅙, ⅓, ⅔, for zero (SEZ) and »big« (SEB) strata effects and for »big« treatment effect (TEB) at $\alpha = 0.05$.

From Table 4 and further simulations by the author it can be concluded that even in situations where alignment provided the highest gains in power with complete observations, it should not be used to analyze trials with more than 33 % censoring. With increasing censoring, the power of alignment tests decreases most rapidly, followed by simple stratification tests and finally by unstratified tests. Therefore alignment will not be considered generally for analyses of censored data from clinical trials.

It is unlikely that other means of alignment will produce an improvement – at least alignment by the Kaplan and Meier (1958) -estimated median or by an interval of the true stratum median of the sample have not produced satisfactory results.

To preserve the power with heavy censoring (≥ 66 %), strata should contain at least 12–16 observations when simple stratification tests are applied.

Because of the independence of the censoring mechanism and because no ties were simulated, the results given for GEHAN's test will be almost the same for BRESLOW's (1974) test, which uses an identical test statistic.

4. Concluding remarks

The decision to stratify in a statistical analysis is taken to increase the power and reduce the bias. Though the latter objective is achieved by both stratified tests, neither is uniformly more powerful, as shown in chapter 3. Alignment should be preferred in studies with balanced or unbalanced strata and »small« strata sizes (possibly $\overline{m_i} \leq 8$) and where censoring is below 16 %.

Though unbalanced treatment frequencies within strata have not been studied by the author, it can be assumed that the reduction in power will affect both stratified tests similarly, and more than the unstratified test. Of course each stratum must contain at least one observation from each treatment.

The three tests compared are all based on U-scores and hence are more powerful to detect translation alternatives. Tests based on exponential scores, such as the tests by SAVAGE (1956) for complete or by MANTEL (1966) for censored observations, would be more powerful in the detection of exponential alternatives.

If strata sizes permit, exploratory tests should check whether treatment effects are similar for the strata considered or whether there are interactions between treatment and strata. Strong interactions may render a combined analysis of the strata meaningless.

Literature

BRESLOW, N. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns od censorship. Biometrika **57,** 579–94.

GEHAN, E. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika **52,** 203–23.

HODGES, J. L., and LEHMANN, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. Ann. Math. Statist. **33,** 482–97.

KAPLAN, E. L., and MEIER, P. (1958). Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. **53,** 457–81.

LEHMANN, E. (1975). Nonparametrics: Statistical Methods based on Ranks. San Francisco: Holden-Day.

LININGER, L., GAIL, M. H., GREEN, S. B., and BYAR, D. P. (1979). Comparison of four tests for equality of survival curves in the presence of stratification and censoring. Biometrika **66,** 419–28.

MANN, H. B.,and WHITNEY, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statist. **18,** 50–60.

MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Rep. **50,** 163–70.

MEHRA, K. L., and SARANGI, J. (1967). Asymptotic efficiency of certain rank tests for comparative experiments. Ann. Math. Statist. **38,** 90–107.

SARANGI, J., and MEHRA, K. L. (1969). Some further results on Hodges-Lehmann conditional rank tests. Bull. Calcutta Statist. Assoc. **18,** 25–41.

SAVAGE, I. R. (1956). Contribution to the theory of rank order statistics – the two sample case. Ann. Math. Statist. **27,** 590–615.

TARDIF, S. (1981). On the almost sure convergence of the permutation distribution for aligned rank test statistics in randomized block designs. Ann. Statist. **9,** 190–93.

Eingegangen am 26. August 1985

Anschrift des Verfassers: Univ.-Doz. Dr. M. Schemper, Arbeitsgruppe Biometrie und Dokumentation, I. Chirurgische Univ.-Klinik, Alser Straße 4, 1090 Wien.

EDV in Medizin und Biologie **16** (4), 121–127, ISSN 0300-8282

Zur Interpretation des multiplen Scheffé-Tests für paarweise Mittelwertvergleiche

Hanspeter Thöni

Zusammenfassung

Die Aussagen von multiplen Mittelwertvergleichen zwischen p Prüfgliedern werden der Betrachtung des gemeinsamen Konfidenzbereiches von $p-1$ linear unabhängigen Prüfglied-Differenzen gegenübergestellt. An einem Beispiel wird gezeigt, wie dieser gemeinsame Konfidenzbereich aussieht und welche Schlüsse über die Prüfglied-Differenzen auf Grund des gemeinsamen Konfidenzbereiches gezogen werden können.

Summary

The conclusions drawn from a multiple comparison procedure applied on p treatment means are contrasted to the analysis of the joint confidence region for $p-1$ linearly independent treatment differences. An example shows the nature of this region, and is used to illustrate the conclusions about treatment differences that can be drawn from looking at the joint confidence region.

1. Einleitung

Bei der Auswertung von Versuchen mit mehr als zwei Prüfgliedern (Behandlungen) stellt sich meist die Frage nach der Beurteilung der paarweisen Mittelwertdifferenzen: Welche Prüfglieder weichen von welchen anderen »signifikant« ab? Zur Beantwortung dieser Frage verwendet man eines der zahlreichen Prüfverfahren für multiple Mittelwertvergleiche [1], [2], [3]. Als Ergebnis erhält man möglicherweise sich gegenseitig überlappende Gruppen von Prüfgliedern, die innerhalb solcher Gruppen als voneinander »nicht signifikant« verschieden, von allen Prüfgliedern außerhalb der betreffenden Gruppe jedoch als »signifikant« verschieden anzusprechen sind. Überlappen sich zwei derartige Gruppen, d. h., gibt es Prüfglieder, die zwei solchen Gruppen angehören, so entsteht die nicht einfach zu interpretierende Situation, daß ein (oder mehrere) Prüfglied(er) von einem Teil unter sich »nicht signifikanter« Prüfglieder ebenfalls »nicht signifikant«, vom Rest derselben Gruppe jedoch »signifikant« verschieden ist (bzw. sind). Da »nicht signifikant« oft interpretiert wird als »Gleichheit der Mittelwerte« der Grundgesamtheiten, so entsteht die widersprüchliche Situation, daß der Erwartungswert einer Grundgesamtheit mit einem Teil unter sich gleicher Erwartungswerte übereinstimmt, vom Rest jedoch signifikant abweicht.

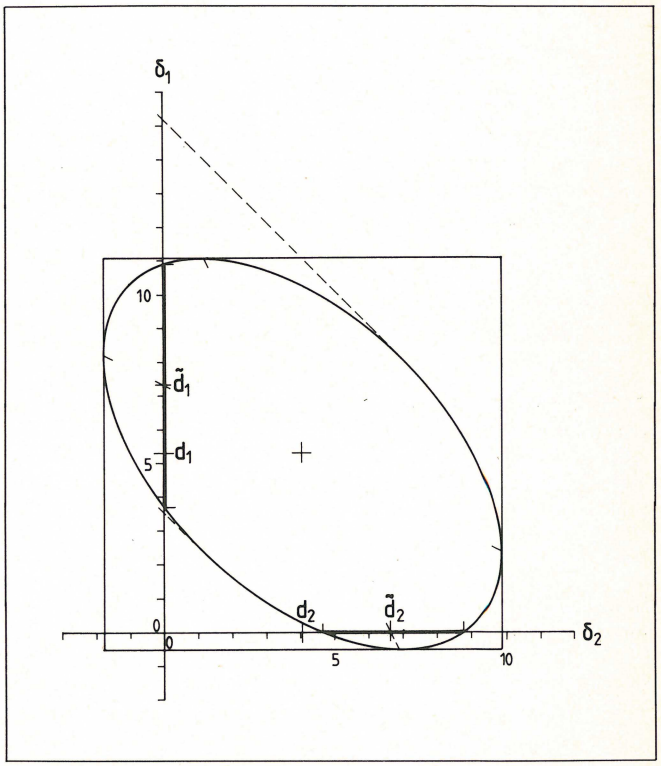


Abb. 1. 0.95-Konfidenzbereich für die Prüfglied-Differenzen δ_1 und δ_2 (vgl. Beispiel 1).

Eine widerspruchsfreie und anschauliche Lösung dieses Problems erhält man durch die Betrachtung des dem SCHEFFÉ-Test zugrundeliegenden $p-1$-dimensionalen gemeinsamen Konfidenzbereiches für $p-1$ linear unabhängige Prüfglied-Differenzen. In Kapitel 2 wird zunächst der einfachste Fall von drei Prüfgliedern und gleich großen Versuchsgruppen behandelt (balancierte Daten). Anhand eines Beispiels wird in Kapitel 3 erläutert, wie die Situation von zwei sich überlappenden »nicht signifikanten«« Prüfglied-Differenzen zu interpretieren ist. In Kapitel 4 erfolgt die Verallgemeinerung auf mehr als drei Prüfglieder und ungleich große Versuchsgruppen (unbalancierte Daten), und in Kapitel 5 wird an einem Beispiel mit sechs Prüfgliedern untersucht, welche der Prüfglied-Differenzen simultan gleich Null sein können.

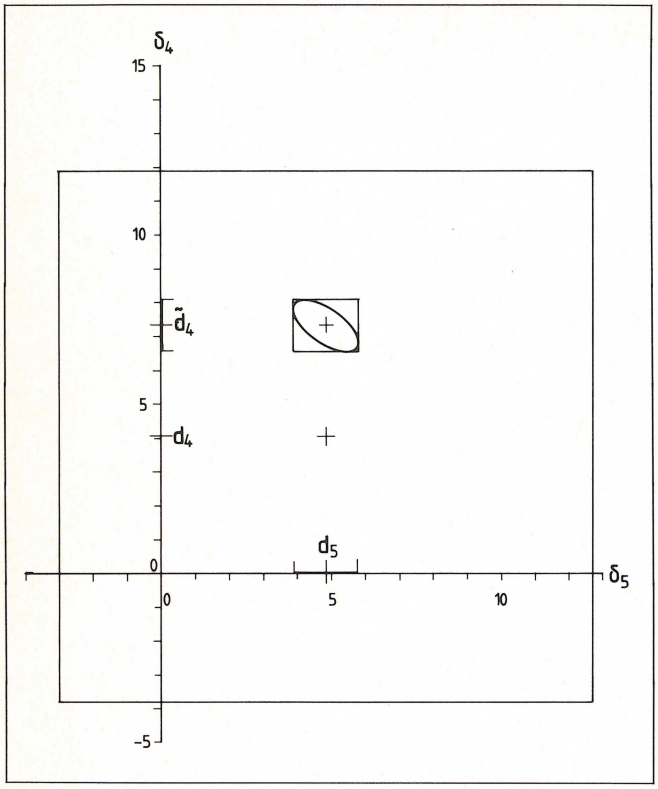


Abb. 2. Bedingter 0.95-Konfidenzbereich für die Prüfglied-Differenzen δ_4 und δ_5 (vgl. Beispiel 2, Abschn. **5.2**).

2. Der gemeinsame Konfidenzbereich für Differenzen zwischen drei Prüfgliedern

Wir gehen aus vom üblichen linearen Varianzanalysemodell

$$y_{ij} = \mu_i + e_{ij} = \mu + \tau_i + e_{ij} \qquad (1)$$

$i = 1, 2, 3$; $j = 1, \ldots, n$. μ_i ist der Erwartungswert der Beobachtungen in der i-ten Versuchsgruppe, μ die gemeinsame Konstante, τ_i der i-te Prüfgliedeffekt, die e_{ij} seien unabhängig identisch normalverteilt mit Erwartungswert Null und Varianz σ^2.

Bei der Versuchsauswertung interessiert zunächst die globale Nullhypothese

$$H_0 : \mu_1 = \mu_2 = \mu_3 \qquad (2)$$

was gleichbedeutend ist mit

$$H_0 : \tau_1 = \tau_2 = \tau_3 = 0 \qquad (3)$$

Die Prüfung dieser globalen Nullhypothese geschieht über die Quadratsummenzerlegung mit dem üblichen F-Test. Wird diese Nullhypothese verworfen, so prüft man mittels eines geeigneten multiplen Testverfahrens, welches der Paare μ_i, $\mu_{i'}$ möglicherweise ungleich ist, d. h., welche Differenz $\tau_i - \tau_{i'} \neq 0$ ist.

Verwenden wir das Testverfahren von SCHEFFÉ [2], so lautet die Prüfvorschrift: verwirf H_0^*: $\tau_i - \tau_{i'} = 0$, falls

$$|\bar{y}_i - \bar{y}_{i'}| > \sqrt{2\, F_{\alpha;2,\nu}\, s^2 \frac{2}{n}}, \qquad i \neq i' \qquad (4)$$

andernfalls verwirf H_0^* nicht. s^2 ist das Mittelquadrat des Versuchsfehlers aus der Quadratsummenzerlegung, $F_{\alpha;2,\nu}$ das α-Quantil der F-Verteilung mit 2 und $\nu = 3(n-1)$ Freiheitsgraden, n der Umfang der Versuchsgruppen.

Die Prüfgröße des SCHEFFÉ-Tests kann umgeformt werden in ein $(1-\alpha)$-Konfidenzintervall für Prüfglied-Differenzen. Es ist $\mu_i - \mu_{i'} = \tau_i - \tau_{i'}$ und

$$\Pr\{\tau_i - \tau_{i'} \in \left((\bar{y}_i - \bar{y}_{i'}) \mp \sqrt{2\, F_{\alpha;2,\nu}\, s^2 \frac{2}{n}}\right)\} \geq 1 - \alpha \qquad (5)$$

das zugehörige Konfidenzintervall.

Bei drei Prüfgliedern können drei derartige Differenzen betrachtet werden, z. B.

$$\begin{aligned} \delta_1 &= \tau_2 - \tau_1 \\ \delta_2 &= \tau_3 - \tau_2 \\ \delta_3 &= \tau_3 - \tau_1 = \delta_1 + \delta_2 \end{aligned} \qquad (6)$$

Die drei Differenzen sind nicht voneinander linear unabhängig, jede kann als Linearkombination der beiden anderen gebildet werden. Für die Berechnung des gemeinsamen Konfidenzbereiches genügt es, diesen für die beiden ersten Differenzen zu berechnen. Der Konfidenzbereich für die dritte ergibt sich aus dem Konfidenzbereich für die Summe der beiden ersten.

Der gemeinsame $(1-\alpha)$-Konfidenzbereich für die Differenzen δ_1 und δ_2 läßt sich wie folgt herleiten. Mit $d_i = \bar{y}_{i+1} - \bar{y}_i$ bezeichnen wir die Schätzfunktionen für die Prüfglied-Differenzen δ_i $(i = 1, 2)$ von (6). Den Vektor der beiden linear unabhängigen Differenzen d_1 und d_2 schreiben wir in Matrizenform

$$\vec{d} = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{pmatrix} \qquad (7)$$

Da die Prüfglied-Mittelwerte voneinander unabhängig sind, ist ihre Kovarianzmatrix eine Diagonalmatrix mit den Elementen σ^2/n. Die Kovarianzmatrix des Vektors $\vec{d}$ lautet sodann

$$V(\vec{d}) = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \sigma^2/n & & 0 \\ & \sigma^2/n & \\ 0 & & \sigma^2/n \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{pmatrix} \qquad (8a)$$

$$= \frac{2\sigma^2}{n} \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix} \qquad (8b)$$

Der **gemeinsame** $(1-\alpha)$-Konfidenzbereich für die Prüfgliedifferenzen δ_1 und δ_2 lautet

$$\left(\frac{2s^2}{n}\right)^{-1} (d_1 - \delta_1,\, d_2 - \delta_2) \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}^{-1} \begin{pmatrix} d_1 - \delta_1 \\ d_2 - \delta_2 \end{pmatrix} - 2F_{\alpha;2,\nu} \leq 0 \qquad (9a)$$

was wir auch in der Form schreiben können

$$(d_1 - \delta_1,\, d_2 - \delta_2) \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}^{-1} \begin{pmatrix} d_1 - \delta_1 \\ d_2 - \delta_2 \end{pmatrix} - 2\, F_{\alpha;2,\nu}\, \frac{2s^2}{n} \leq 0 \qquad (9b)$$

Setzen wir in (9b) das Gleichheitszeichen, so erhalten wir die Gleichung einer Ellipse mit Zentrum (d_1, d_2), deren Achsen gegen die δ_1- bzw. δ_2-Achsen um $-45°$ geneigt sind. Die Halbachsen haben die Länge

$$\sqrt{3}\sqrt{2F_{\alpha;2,\nu}\frac{s^2}{n}},\ \sqrt{2F_{\alpha;2,\nu}\frac{s^2}{n}} \qquad (10)$$

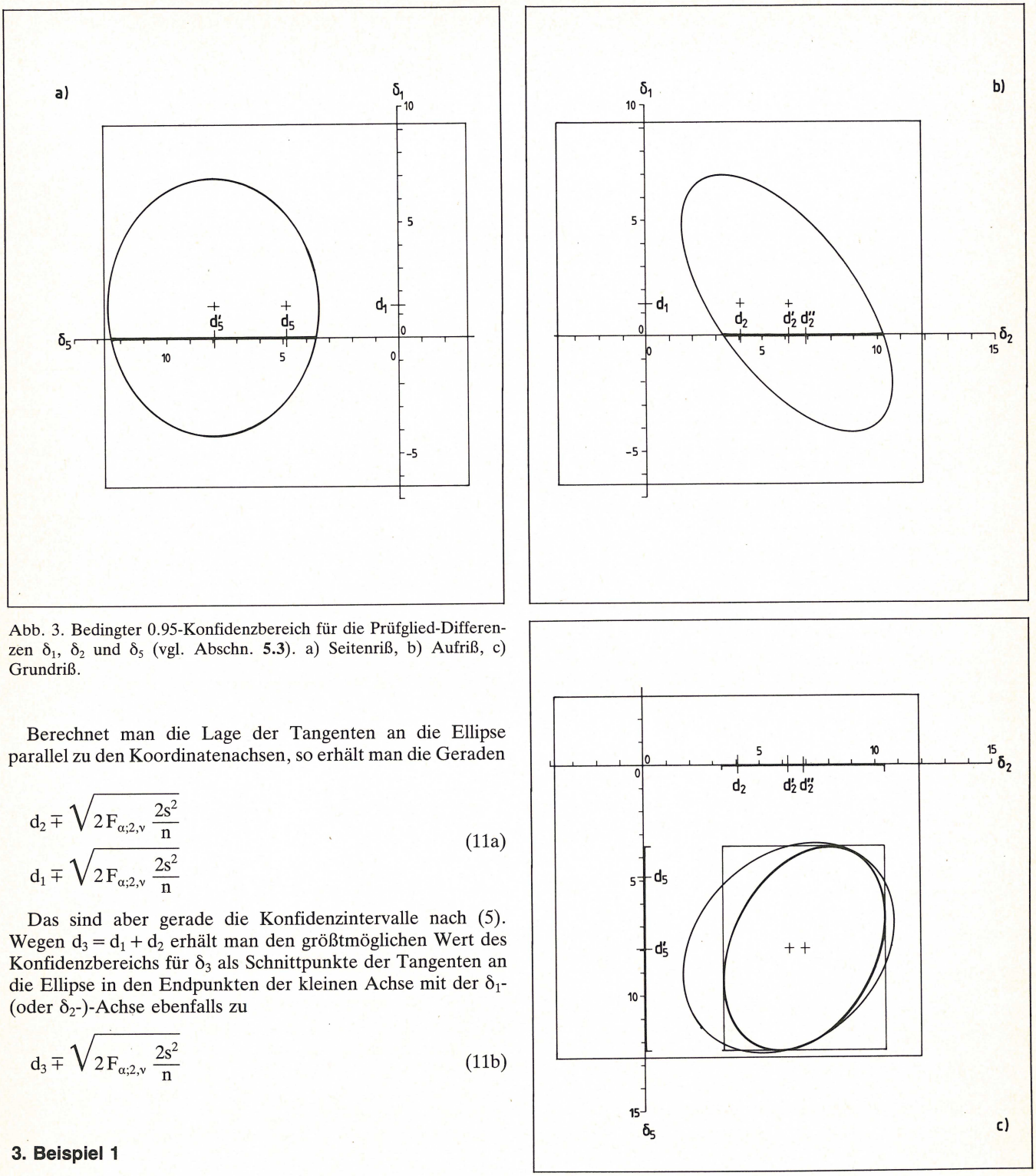


Abb. 3. Bedingter 0.95-Konfidenzbereich für die Prüfglied-Differenzen δ_1, δ_2 und δ_5 (vgl. Abschn. **5.3**). a) Seitenriß, b) Aufriß, c) Grundriß.

Berechnet man die Lage der Tangenten an die Ellipse parallel zu den Koordinatenachsen, so erhält man die Geraden

$$d_2 \mp \sqrt{2F_{\alpha;2,\nu}\frac{2s^2}{n}}$$
$$d_1 \mp \sqrt{2F_{\alpha;2,\nu}\frac{2s^2}{n}} \qquad (11a)$$

Das sind aber gerade die Konfidenzintervalle nach (5). Wegen $d_3 = d_1 + d_2$ erhält man den größtmöglichen Wert des Konfidenzbereichs für δ_3 als Schnittpunkte der Tangenten an die Ellipse in den Endpunkten der kleinen Achse mit der δ_1- (oder δ_2-)-Achse ebenfalls zu

$$d_3 \mp \sqrt{2F_{\alpha;2,\nu}\frac{2s^2}{n}} \qquad (11b)$$

3. Beispiel 1

STEEL and TORRIE [4] geben die Ergebnisse eines Versuchs zur Stickstoffbindung durch Knöllchenbakterien bei Rotklee wieder. Wir verwenden zunächst drei der insgesamt sechs Prüfglieder zur Illustration der Ergebnisse in Kapitel 2. Die Quadratsummenzerlegung zu den Daten in Tabelle 1 lautet:

	SQ	FG	MQ
zwischen d. Prüfgliedern	219.329	2	109.665
Versuchsfehler	129.948	12	10.829

Mit $F = 10.13$, $F_{0.05;2,12} = 3.89$ kann die globale Nullhypothese verworfen werden. Der SCHEFFÉ-Test ergibt, daß Prüfglied 1 von Prüfglied 2 und Prüfglied 2 von Prüfglied 3 nicht signifikant verschieden ist, jedoch Prüfglied 1 von Prüfglied 3 signifikant abweicht.

Betrachten wir den gemeinsamen 0.95-Konfidenzbereich der Differenzen δ_1 und δ_2 (vgl. Abb. 1). Die Gleichung der Konfidenzellipse lautet gemäß (9b)

$$(5.28 - \delta_1,\ 4.06 - \delta_2)\begin{pmatrix} 1 & 0.5 \\ -0.5 & 1 \end{pmatrix}^{-1}\begin{pmatrix} 5.28-\delta_1 \\ 4.06-\delta_2 \end{pmatrix} - 33.70 = 0 \qquad (12)$$

Tabelle 1. Stickstoffgehalt in mg von Rotklee-Pflanzen, die mit drei verschiedenen Rhizobium-Stämmen beimpft wurden.
([4], Tabelle 7.1, Seite 140)

Rhizobium-Stamm	3D0k4	3D0k7	3D0k5
	17.0	20.7	17.7
	19.4	21.0	24.8
	9.1	20.5	27.9
	11.9	18.8	25.2
	15.8	18.6	24.3
$\bar{y}_i$	14.64	19.92	23.98

$d_1 = 5.28$ $\quad d_2 = 4.06$
$d_3 = 9.34$

Die Ellipse überschneidet sowohl die δ_2-Achse ($\delta_1 = 0$) als auch die δ_1-Achse ($\delta_2 = 0$), überdeckt jedoch den Punkt $\delta_1 = 0$, $\delta_2 = 0$ nicht. Das bedeutet, daß entweder $\delta_1 = 0$ (bzw. $\mu_2 - \mu_1 = 0$) **oder** $\delta_2 = 0$ (bzw. $\mu_3 - \mu_2 = 0$) sein kann, nicht aber beide Aussagen gleichzeitig richtig sein können. Obwohl die einfachen Konfidenzintervalle für

$$\mu_2 - \mu_1 = \delta_1 \in (-0.53,\ 11.09]$$

und (13a)

$$\mu_3 - \mu_2 = \delta_2 \in (-1.75,\ 9.87]$$

den Wert 0 jeweils einschließen, verläuft die Gerade $\delta_3 = \delta_1 + \delta_2 = 0$ ganz außerhalb der Ellipse, und das zugehörige Konfidenzintervall

$$\mu_3 - \mu_1 = \delta_3 \in (3.53,\ 15.15] \qquad (13b)$$

schließt den Wert 0 nicht ein. Aus der Gleichung für die Konfidenzellipse läßt sich berechnen, in welchem Bereich z. B. δ_2 mit Wahrscheinlichkeit $1-\alpha = 0.95$ liegt, wenn $\delta_1 = 0$ angenommen wird (bedingter Konfidenzbereich). Man erhält das Intervall

$$4.61 < \delta_2 \leqslant 8.79,\ \delta_1 = 0$$

bzw.

$$3.71 < \delta_1 \leqslant 10.91,\ \delta_2 = 0$$

Die Punkte

$$\bar{y}_3 - \tfrac{1}{2}(\bar{y}_1 + \bar{y}_2) = 6.70 = \tilde{d}_2$$

bzw.

$$\tfrac{1}{2}(\bar{y}_2 + \bar{y}_3) - \bar{y}_1 = 7.31 = \tilde{d}_1$$

halbieren diese bedingten Konfidenzintervalle.

(Einzelheiten zur Darstellung der Ellipse und deren Tangenten vgl. z. B. [5].)

4. Verallgemeinerung auf p > 3 Prüfglieder

Die Ergebnisse aus Kapitel 2 lassen sich leicht auf mehr als drei Prüfglieder und ungleich große Versuchsgruppen verallgemeinern. Unter den $\binom{p}{2}$ möglichen Paardifferenzen zwischen den p Prüfgliedern gibt es genau $p-1$ voneinander linear unabhängige, z. B. die $p-1$ Prüfglied-Differenzen $\delta_i = \tau_{i+1} - \tau_i$ ($i = 1, 2, \ldots, p-1$). Die restlichen $(p-1)(p-2)/2$ können als Linearkombinationen dieser Differenzen gebildet werden.

Die Schätzfunktionen $d_i = \bar{y}_{i+1} - \bar{y}_i$ können wiederum in der Form

$$\vec{d} = \begin{pmatrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ \cdot \\ d_{p-1} \end{pmatrix} = \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & 0 \\ & & \cdot & \cdot & \\ & & & \cdot & \cdot \\ 0 & & & -1 & 1 \end{pmatrix} \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \cdot \\ \cdot \\ \cdot \\ \bar{y}_p \end{pmatrix} \qquad (14)$$

geschrieben werden. Die Kovarianzmatrix der Prüfglied-Mittelwerte ist eine Diagonalmatrix mit den Elementen σ^2/n_i, so daß die Kovarianzmatrix der Prüfglied-Differenzen die Form

$$V(\vec{d}) = \sigma^2 \begin{bmatrix} (\frac{1}{n_1} + \frac{1}{n_2}) & -\frac{1}{n_2} & & & 0 \\ -\frac{1}{n_2} & (\frac{1}{n_2} + \frac{1}{n_3}) & -\frac{1}{n_3} & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot \\ & & -\frac{1}{n_{p-2}} & (\frac{1}{n_{p-2}} + \frac{1}{n_{p-1}}) & -\frac{1}{n_{p-1}} \\ 0 & & & -\frac{1}{n_{p-1}} & (\frac{1}{n_{p-1}} + \frac{1}{n_p}) \end{bmatrix} \qquad (15)$$

$$= \sigma^2 U$$

erhält. Für balancierte Daten ($n_i = n\ \forall\ i$) vereinfacht sich (15) zu

$$V(\vec{d}) = \frac{2\sigma^2}{n} \begin{pmatrix} 1 & -0.5 & & & \\ -0.5 & 1 & -0.5 & & 0 \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot \\ & & & -0.5 & 1 & -0.5 \\ 0 & & & & -0.5 & 1 \end{pmatrix} \qquad (16)$$

$$= \frac{2\sigma^2}{n} R.$$

Ersetzt man wiederum σ^2 durch die Schätzung s^2 aus der Quadratsummenzerlegung mit $\nu = N-p$ Freiheitsgraden, so lautet die Gleichung des $(p-1)$-dimensionalen $(1-\alpha)$-Konfidenzbereiches

$$(\vec{d} - \vec{\delta})' U^{-1} (\vec{d} - \vec{\delta}) - (p-1) F_{\alpha;p-1,\nu} s^2 \leqslant 0 \qquad (17a)$$

und vereinfacht sich im Falle balancierter Daten zu

$$(\vec{d} - \vec{\delta})' R^{-1} (\vec{d} - \vec{\delta}) - (p-1) F_{\alpha;p-1,\nu} \frac{2s^2}{n} \leqslant 0 \qquad (17b)$$

Setzt man in (17) das Gleichheitszeichen, so erhält man die Gleichung eines $(p-1)$-dimensionalen Ellipsoids, Alle Vektoren $\vec{\delta}$, welche (17) erfüllen, liegen im Innern des gemeinsamen $(1-\alpha)$-Konfidenzbereiches. Die Hyperebenen

$$d_i \pm \sqrt{(p-1) F_{\alpha;p-1,\nu} s^2 (\frac{1}{n_i} + \frac{1}{n_{i+1}})} \qquad (18)$$

bilden die Tangentialebenen an das Ellipsoid senkrecht zu den δ_i-Achsen.

5. Beispiel 2

5.1

Das in Kapitel 3 verwendete Beispiel ist Teil eines umfangreicheren Experiments mit 6 Prüfgliedern (Tabelle 2). Aus den vollständigen Daten errechnet man die folgende Quadratsummenzerlegung

	SQ	FG	MQ
zwischen d. Prüfgliedern	847.047	5	169.409
Versuchsfehler	282.901	24	11.788

und für die Prüfgröße F = 14.37. Mit $F_{0.05;5,24} = 2.62$ kann die globale Nullhypothese verworfen werden; für die halbe Länge der einfachen 0.95-Konfidenzintervalle für die d_i erhalten wir 7.86. Alle Differenzen zwischen zwei Prüfgliedern, die kleiner als 7.86 sind, sind »nicht signifikant« von Null verschieden. Dies trifft für alle Differenzen zwischen aufeinanderfolgenden Prüfgliedern zu, d. h., $\delta_i = 0$ liegt im Innern eines jeden individuellen Konfidenzintervalls.

Die Gleichung des fünfdimensionalen 0.95-Konfidenzbereiches lautet

$$(1.38 - \delta_1, \ldots, 4.84 - \delta_5)$$

$$\begin{pmatrix} 1 & -0.5 & & & \\ -0.5 & 1 & -0.5 & 0 & \\ & -0.5 & 1 & -0.5 & \\ & 0 & -0.5 & 1 & -0.5 \\ & & & -0.5 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1.38 - \delta_1 \\ 4.06 - \delta_2 \\ 1.22 - \delta_3 \\ 4.06 - \delta_4 \\ 4.84 - \delta_5 \end{pmatrix}$$

$$-5\,(2.62)\,(2)\,(11.788)/5 \leqslant 0 \qquad (19)$$

Setzt man $\vec{\delta}' = (0, 0, 0, 0, 0)$ in die Gleichung (19) ein, so erhält man $338.819 - 61.769 > 0$, d. h., der Nullvektor liegt außerhalb des gemeinsamen 0.95-Konfidenzbereichs, obwohl jede einzelne Differenz $\delta_i = 0$ innerhalb des individuellen Konfidenzintervalls liegt.

Mit anderen Worten: Es können nicht alle fünf δ_i **gleichzeitig** Null sein (was natürlich mit der Aussage des globalen F-Tests übereinstimmen muß!).

Die Anwendung des SCHEFFÉ-Tests ergibt, daß die Differenzen $\bar{y}_4 - \bar{y}_1 = 6.66$, $\bar{y}_5 - \bar{y}_3 = 5.28$ und $\bar{y}_6 - \bar{y}_5 = 4.84$ kleiner sind als die kritische Differenz 7.86. In der herkömmlichen Weise können somit drei sich überlappende »nicht signifikante« Gruppen von Prüfgliedern gebildet werden. Wie verhält es sich mit der Gleichzeitigkeit?

5.2

Die Prüfglieder 1, 2, 3, 4 bilden eine »nicht signifikante« Gruppe. Es liegt nun nahe, den bedingten Konfidenzbereich für δ_4 und δ_5 zu bestimmen unter der Einschränkung $\delta_1 = \delta_2 = \delta_3 = 0$. In (19) eingesetzt erhalten wir

$$8\,\delta_4^2 + 8\,\delta_4\delta_5 + 5\,\delta_5^2 - 2(78.16)\delta_4 - 2(53.60)\delta_5 + 831.149 \leqslant 0 \qquad (20)$$

Mit dem Gleichheitszeichen erhalten wir wiederum die Gleichung einer Ellipse mit Zentrum $\tilde{d}_4 = 7.35$, $d_5 = 4.84$ und

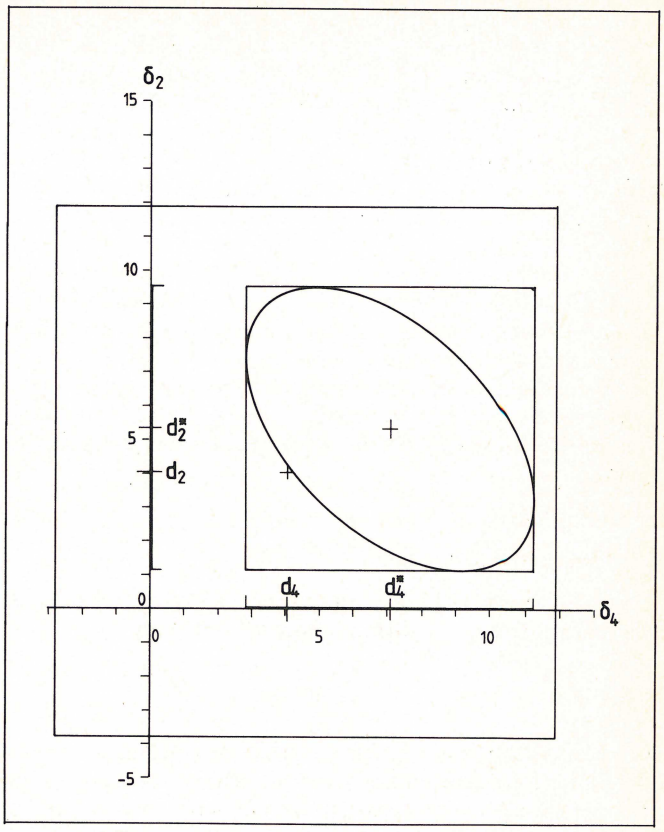


Abb. 4. Bedingter 0.95-Konfidenzbereich für die Prüfglied-Differenzen δ_2 und δ_4 (vgl. Abschn. **5.4**).

um −34.7° geneigten Achsen der Länge 1.11 bzw. 0.51. Aus der Abbildung 2 ersieht man, daß diese Ellipse vollständig innerhalb des Quadrates liegt, das von den Geraden $\delta_4 = 4.06 \mp 7.86$ und $\delta_5 = 4.84 \mp 7.86$ begrenzt wird, und die δ_4- bzw. δ_5-Achse nirgendwo überschneidet. Dies bedeutet, daß, wenn wir annehmen, die Erwartungswerte der Prüfglieder 1, 2, 3, 4 stimmten überein, die Differenzen δ_4 und δ_5 nicht gleichzeitig auch Null sein können:

$$\delta_1 = \delta_2 = \delta_3 = 0;\ \delta_4 \neq 0,\ \delta_5 \neq 0$$

Der Wert für $\tilde{d}_4$ stimmt überein mit der Differenz zwischen dem arithmetischen Mittel der ersten vier und dem fünften Prüfglied:

$$\bar{y}_{1234} = 16.63;\ \bar{y}_5 - \bar{y}_{1234} = 7.35 = \tilde{d}_4$$

Tabelle 2. Stickstoffgehalt von Rotklee-Pflanzen, die mit verschiedenen Rhizobium-Stämmen bzw. einer Mischkultur beimpft wurden ([4], Tabelle 7.1, Seite 140).

Rhizobium-Stamm	3D0k13	3D0k4	Mischkultur	3D0k7	3D0k5	3D0k1
	14.3	17.0	17.3	20.7	17.7	19.4
	14.4	19.4	19.4	21.0	24.8	32.6
	11.8	9.1	19.1	20.5	27.9	27.0
	11.6	11.9	16.9	18.8	25.2	32.1
	14.2	15.8	20.8	18.6	24.3	33.0
$\bar{y}_i$	13.26	14.64	18.70	19.92	23.98	28.82
d_i		1.38	4.06	1.22	4.06	4.84

Die horizontalen Tangenten an die Ellipse schließen einen Konfidenzbereich für $\bar{\delta}_4 \in (6.59, 8.11]$ ein. $\bar{d}_5$ stimmt mit d_5 überein, die senkrechten Tangenten an die Ellipse begrenzen den Konfidenzbereich für $\delta_5 \in (3.88, 5.80]$. Daß das Intervall für $\bar{\delta}_4$ kürzer ausfällt als dasjenige für δ_5, erklärt sich daraus, daß $\bar{y}_{1234}$ das arithmetische Mittel aus 4n Beobachtungen ist und die Varianz von $\bar{d}_4$ $5\sigma^2/4n$ beträgt, diejenige von d_5 jedoch $2\sigma^2/n$. Die Längen der beiden Intervalle verhalten sich wie $\sqrt{5/8} = 0.791$.

Die sich aus den beiden aneinanderstoßenden »nicht signifikanten« Prüfgliedgruppen 1, 2, 3, 4 und 5, 6 **scheinbar** ergebende Schlußfolgerung, die sechs Prüfglieder ließen sich in zwei unter sich homogene Gruppen mit $\mu_1 = \mu_2 = \mu_3 = \mu_4$ und $\mu_5 = \mu_6$ zerlegen, ist somit **nicht** erlaubt! Dies erkennt man sofort daran, daß der Vektor $\bar{\delta}' = (0, 0, 0, \delta_4, 0)$ **nicht** im Innern des Konfidenzbereiches liegt; die quadratische Gleichung

$$8\,\delta_4^2 - 156.32\,\delta_4 + 831.149 = 0 \qquad (21)$$

hat keine reellen Wurzeln.

Für die beiden Prüfgliedgruppen 1, 2, 3, 4 und 5, 6 kann also die Annahme $\mu_i = \mu_{i'}$ **nicht gleichzeitig** zutreffen.

5.3

In derselben Weise können wir auch den bedingten Konfidenzbereich betrachten unter der Annahme $\mu_3 = \mu_4 = \mu_5$ (vgl. Tabelle 2), d. h., wir setzen $\delta_3 = \delta_4 = 0$. Setzen wir dies in (19) ein, so resultiert daraus der bedingte 0.95-Konfidenzbereich für δ_1, δ_2 und δ_5. Die Gleichung des den Konfidenzbereich begrenzenden Ellipsoids lautet

$$(\delta_1, \delta_2, \delta_5, 1) \begin{pmatrix} 5 & 4 & 1 & -39.76 \\ 4 & 8 & 2 & -71.24 \\ 1 & 2 & 5 & -53.60 \\ -39.76 & -71.24 & -53.60 & 831.149 \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_5 \\ 1 \end{pmatrix} = 0 \qquad (22)$$

Sein Zentrum hat die Koordinaten

$$\begin{aligned} d_1 &= \bar{y}_2 - \bar{y}_1 = 1.38 \\ d_2' &= \frac{1}{3}(\bar{y}_3 + \bar{y}_4 + \bar{y}_5) - \bar{y}_2 = 6.227 \\ d_5' &= \bar{y}_6 - \frac{1}{3}(\bar{y}_3 + \bar{y}_4 + \bar{y}_5) = 7.953 \end{aligned} \qquad (23)$$

Abbildung 3 (siehe S. 123) zeigt die Projektionen dieses Ellipsoids auf die δ_1,δ_2-, δ_1,δ_5- und δ_2,δ_5-Ebene (Seitenriß, Aufriß, Grundriß, vgl. z.B. [6]). Man sieht, daß es die δ_2,δ_5-Ebene ($\delta_1 = 0$) durchstößt, nicht jedoch die Ebenen $\delta_2 = 0$ und $\delta_5 = 0$. Die Schnittfigur ist die Ellipse

$$(\delta_2, \delta_5, 1) \begin{pmatrix} 8 & 2 & -71.24 \\ 2 & 5 & -53.60 \\ -71.24 & -53.60 & 831.149 \end{pmatrix} \begin{pmatrix} \delta_2 \\ \delta_5 \\ 1 \end{pmatrix} = 0 \qquad (24)$$

mit dem Zentrum

$$\begin{aligned} d_2'' &= \frac{1}{3}(\bar{y}_3 + \bar{y}_4 + \bar{y}_5) - \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = 6.917 \\ d_5' &= 7.953 \end{aligned} \qquad (25)$$

Der Punkt $\delta_1 = \delta_3 = \delta_4 = 0$ liegt somit im Innern des 0.95-Konfidenzbereiches; unter dieser Einschränkung liegen δ_2 und δ_5 innerhalb der Ellipse (24) (vgl. Abb. 3c), deren vertikale und horizontale Tangenten die Intervalle

$$d_2'' \mp 3.50 = (3.42, 10.41]$$

und

$$d_5' \mp 4.42 = (3.53, 12.37]$$

begrenzen. Die Längen dieser Intervalle verhalten sich wie die Quadratwurzel aus dem Quotienten der Varianzen von d_2'' und d_5':

$$\sqrt{\frac{5}{6}/\frac{4}{3}} = 0.791.$$

5.4

Als dritte Möglichkeit betrachten wir den Fall $\delta_5 = 0$. Der vierdimensionale bedingte Konfidenzbereich schließt die Werte $\delta_1 = 0$ und $\delta_3 = 0$ ein; der gemeinsame bedingte Konfidenzbereich für δ_2 und δ_4 erhält die Form

$$(\delta_2 - 5.36, \delta_4 - 7.09) \begin{pmatrix} 8 & 4 \\ 4 & 8 \end{pmatrix} \begin{pmatrix} \delta_2 - 5.36 \\ \delta_4 - 7.09 \end{pmatrix} - 104.8558 \leqslant 0 \qquad (26)$$

Das Zentrum der Ellipse stimmt überein mit den Differenzen

$$\begin{aligned} d_2^* &= \frac{1}{2}(\bar{y}_3 + \bar{y}_4) - \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = 5.36 \\ \text{und} \qquad & \qquad (27) \\ d_4^* &= \frac{1}{2}(\bar{y}_6 + \bar{y}_5) - \frac{1}{2}(\bar{y}_3 + \bar{y}_4) = 7.09 \end{aligned}$$

die horizontalen und vertikalen Tangenten begrenzen den Bereich

$$\begin{aligned} & d_2^* \mp 4.18 = (1.18, 9.54] \\ \text{und} \qquad & \qquad (28) \\ & d_4^* \mp 4.18 = (2.91, 11.27] \end{aligned}$$

und schließen den Punkt $\delta_2 = \delta_4 = 0$ nicht ein (vgl. Abb. 4). Wenn wir also annehmen, daß die Prüfglieder 5 und 6 übereinstimmende Mittelwerte aufweisen, können die Erwartungswerte der Prüfglieder 1–4 nicht auch gleichzeitig alle übereinstimmen (man vergleiche diesen Befund mit dem Ergebnis in 5.2!).

5.5

Zusammenfassend kann die Analyse des gemeinsamen 0.95-Konfidenzbereiches für die Prüfglied-Differenzen $\delta_1 = \mu_2 - \mu_1$, $\delta_2 = \mu_3 - \mu_2$, . . ., $\delta_5 = \mu_6 - \mu_5$ wie folgt beschrieben werden:

Der gemeinsame 0.95-Konfidenzbereich wird begrenzt durch das Ellipsoid (19). Er schließt sowohl den Nullvektor $(\delta_1, \delta_2, \delta_3) = \vec{0}'$ als auch $(\delta_1, \delta_3, \delta_4) = \vec{0}'$ und $(\delta_1, \delta_3, \delta_5) = \vec{0}'$ ein. Die **bedingten** 0.95-Konfidenzbereiche für die verbleibenden, von Null verschiedenen Prüfglieddifferenzen werden durch die Ellipsen (20) (für δ_4 und δ_5, vgl. Abb. 2), (24) (für δ_2 und δ_4, vgl. Abb. 3c) bzw. (26) (für δ_2 und δ_4, vgl. Abb. 4) begrenzt. Die Werte $\delta_1 = 0$ und $\delta_3 = 0$ liegen in jedem Fall innerhalb des Konfidenzbereiches, von den Differenzen δ_2, δ_4 und δ_5 liegt immer nur für **eine** Differenz der Wert Null innerhalb des Konfidenzbereiches; die bedingten Konfidenzbereiche für die beiden anderen schließen dann aber den Wert Null aus! Die Versuchsergebnisse lassen somit die folgenden, sich **gegenseitig ausschließenden** Interpretationen der Prüfglied-Mittelwerte und -Differenzen zu:

5.2) $\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}_3 = \hat{\mu}_4 = 16.63$
$\hat{\mu}_5 = 23.98$
$\hat{\mu}_6 = 28.82$
$6.59 < \hat{\mu}_5 - \hat{\mu}_4 \leq 8.11$
$3.88 < \hat{\mu}_6 - \hat{\mu}_5 \leq 5.80$

5.3) $\hat{\mu}_1 = \hat{\mu}_2 = 13.95$
$\hat{\mu}_3 = \hat{\mu}_4 = \hat{\mu}_5 = 20.87$
$\hat{\mu}_6 = 28.82$
$3.42 < \hat{\mu}_3 - \hat{\mu}_2 \leq 10.41$
$3.53 < \hat{\mu}_6 - \hat{\mu}_5 \leq 12.37$

5.4) $\hat{\mu}_1 = \hat{\mu}_2 = 13.95$
$\hat{\mu}_3 = \hat{\mu}_4 = 19.31$
$\hat{\mu}_5 = \hat{\mu}_6 = 26.40$
$1.18 < \hat{\mu}_3 - \hat{\mu}_2 \leq 9.54$
$2.91 < \hat{\mu}_5 - \hat{\mu}_4 \leq 11.27$

Von den auf Grund des SCHEFFÉ-Tests gebildeten »nicht signifikant« verschiedenen Prüfgliedgruppen (1, 2, 3, 4), (3, 4, 5) und (5, 6) kann die Nullhypothese: $\mu_i = \mu_{i'}$ immer **nur für eine** Gruppe gelten, nicht aber beispielsweise für die erste und dritte Gruppe gleichzeitig!

Um zu einer Rangordnung unter den Alternativen in **5.2, 5.3** bzw. **5.4** zu gelangen, berechnen wir den numerischen Wert der linken Seite der Ungleichung (19) für den jeweiligen Vektor $\vec{\delta}$. Mit $\vec{\delta}' = (0, 0, 0, 7.35, 4.84)$ erhalten wir für die quadratische Form $(\vec{d} - \vec{\delta})'R^{-1}(\vec{d} - \vec{\delta})$ den Wert 60.852, $\vec{\delta}' = (0, 6.917, 0, 0, 7.953)$ liefert 32.471 und $\vec{\delta}' = (0, 5.36, 0, 7.09, 0)$ schließlich 26.818. Die linke Seite von (19) ergibt somit -0.917 für die Alternative **5.2**, -29.298 für **5.3** und -34.951 für **5.4.** Je kleiner dieser Wert ausfällt, desto tiefer im Innern des Konfidenzbereiches befindet sich $\vec{\delta}$, d. h., desto größer ist seine Mutmaßlichkeit. Die Berechnungen ergeben somit, daß die Alternative **5.4** den kleinsten Wert für (19) liefert und somit als die wahrscheinlichste unter den drei betrachteten angesehen werden kann.

6. Diskussion

Bei der Interpretation multipler Mittelwertvergleiche ist Vorsicht geboten. Nicht immer trifft es zu, daß Prüfglieder, die auf Grund eines multiplen Mittelwertvergleiches zu »nicht signifikant verschiedenen« oder »homogenen« Gruppen zusammengefaßt werden können, **gleichzeitig** identische Erwartungswerte aufweisen, auch wenn sich diese Gruppen nicht überlappen.

Besser als die Berechnung kritischer Differenzen ist die Betrachtung des **gemeinsamen Konfidenzbereiches** von linear unabhängigen Prüfglied-Differenzen. Die Analyse des diesen Bereich einhüllenden $(p-1)$-dimensionalen Ellipsoids gibt genauen Aufschluß darüber, in welchen Bereichen sich die Prüfglied-Differenzen **gleichzeitig** bewegen können.

Wie anhand des Beispiels 2 gezeigt wird, kann die ausschließliche Betrachtung der »nicht signifikant verschiedenen« Gruppen zu Fehlinterpretationen führen.

Danksagung

Herrn Prof. Dr. E. SONNEMANN, Trier, danke ich für wertvolle Anregungen, die zu einer Verbesserung der ursprünglichen Fassung dieser Arbeit führten.

Literatur

[1] BERCHIER, P. (1981): Mittelwertvergleiche in Normalverteilungsmodellen. In: Simultane Hypothesenprüfungen. Hrsg. von U. FERNER. Biometrisches Seminar der Region Österreich-Schweiz der Internationalen Biometrischen Gesellschaft.
[2] SCHEFFÉ, H. (1959): The Analysis of Variance. Wiley, New York. p. 68 ff.
[3] SONNEMANN, E. (1982): Allgemeine Lösungen multipler Testprobleme. EDV in Medizin und Biologie **13,** 120–128.
[4] STEEL, R. G. D., J. H. TORRIE (1980): Principles and Procedures of Statistics. 2nd edition, McGraw Hill Inc., New York.
[5] THÖNI, H. (1984): Die Berechnung von ausgewählten Eckpunkten einer Konfidenz- bzw. Toleranz-Ellipse. EDV in Medizin und Biologie **15,** 53–57.
[6] BACH, G., H. THÖNI (1981): Darstellung eines Toleranz- und Vertrauensellipsoids. EDV in Medizin und Biologie **12,** 57–61.

Eingegangen am 21. Oktober 1985
Anschrift des Verfassers: Prof. Dr. Hanspeter Thöni, Institut für Angewandte Mathematik und Statistik, Universität Hohenheim, Postfach 70 05 62, D-7000 Stuttgart 70.

EDV in Medizin und Biologie **16** (4), 128–134, ISSN 0300-8282

Computational Methods to Determine a Break Point in Linear Regression

Lutz Edler and Jutta Berger

Summary

A computer program has been implemented for the computational determination of a break point of a linear regression. It comprises several statistical procedures and allows for graphical examination. Its realisation by the APL program CHREG provides an exploratory statistical tool for the analysis of time/dose – response data for deviations from linearity.

Key words: Break Point, Change-Point, Linear Regression, Least Squares, Recursive Residuals, Cumulative Sum (CUSUM)

Zusammenfassung

Für die Bestimmung des Bruchpunkts einer linearen Regression wurde ein APL-Programm entwickelt, welches verschiedene statistische Verfahren anwendet und grafische Bewertungen ermöglicht. Das Programm wurde als APL-Funktion CHREG realisiert und gibt dem Biometriker die Möglichkeit, Zeit-/Dosis-Wirkungs-Daten auf Abweichungen von der Linearitätsannahme explorativ statistisch auszuwerten.

I. Introduction

Specification of statistical models is the most essential and most consequential but also the most difficult step in the analysis of experimental data. Even if a class of models has been identified, there often rest some characteristics of the model to be specified in order to identify the model unambiguously. This can be achieved by theoretical considerations using substantial knowledge from the underlying scientific basis. However, in fundamental research there is often a considerable amount of uncertainty with respect to the underlying mechanisms which forces to look for other solutions. An appealing approach is to let – within a prespecified class of models – the observed data themselves determine the definite model.

This general idea will be applied and exemplified in the following for the problem of fitting a straight line to experimental data (e.g. time series or dose-response relationships) if the extent of the linearity assumption is not known in advance. This applies to a very common situation in applied biostatistics when there is strong evidence for a linear relationship between the independent variable x and the dependent variable y in some range of x values, but when there is also evidently nonlinearity in other ranges, because of the presence of several phases or because of nonlinear behavior at the boundaries. In statistical theory, this problem has found attention as change-point problem and has been treated within the framework of linear regression (HUDSON (1966), BROWN et al. (1975), MCCABE and HARRISON (1980), SCHULZE (1984)) as well as within the more general framework of the change of the distribution in the series of observations y_t (HINKLEY (1970), DESHAYES and PICARD (1980)). Many of the methods were designed for time series with usually a large amount of data. The aspect of shorter series – very common in the case of observations from medical or biological applications – has been neglected mostly and has found almost no regard in standard statistical software. This motivated us to implement computational statistical procedures for the estimation of break points applicable also for shorter series ((x_t, y_t), $t = 1, \ldots, T$).

The following is confined to the simple linear regression and the examination of a break point at the right boundary. This one-sided view covers a large amount of practical problems and in many applications it should be possible to split a two-sided problem into two one-sided. The restriction to one-dimensional regression is more serious, but again one could argue that this simple models occurs in many applications. On the other hand, a generalization to multiple regression can be done for some procedures straightforwardly.

Let us precise the problem:

We assume $x_1 < x_2 <, \ldots, < x_T$ to denote a monotone increasing sequence of independent variables with corresponding values y_t of the dependent variable. The regression model is given by

$$(1) \quad y_t = \begin{cases} b_0 + b_1 x_t + e_t, & t = 1, \ldots, r^* \\ f(x_t) + e_t, & t = r^* + 1, \ldots, T \end{cases}$$

where $f(x_t)$ denotes an unspecified function describing the relation between y and x from x_{r^*+1} on. b_0, b_1 and r^* are unknown parameters. For the error term e_t we assume throughout a normal distribution with homogeneous but unknown variance σ^2. r^* is the parameter to be estimated whereas b_0, b_1 and σ^2 are nuisance parameters. Notice the simplification of the problem by defining the break point in terms of the indices of the data and excluding so a break point x^* in the interval (x_{r^*}, x_{r^*+1}).

A computational solution of this estimation problem was impossible by statistical standard software in a straightforward way. The necessity to program special break point methods anyway – in a command language of a package or in its original language – was taken as opportunity to develop a selfcontained program containing special statistical features of the regression problem.

In the next section we describe the estimation problem of r* if f in (1) is linear. In Section 3 we give a review of the main computational methods for the estimation of r* and their application to model (1). The implementation the APL program CHREG is described in Section 4.

II. Regression Model and Notations

The estimation problem of r* can be formulated in terms of *segmented regression*

$$(2)\quad E[Y] = \begin{cases} b_{01} + b_{11}x & a \le x \le x^* \\ b_{02}x + b_{12}x & x^* \le x \le b. \end{cases}$$

E denotes the conditional expectation of Y given X. Continuity is assumed at x* by $b_{01} + b_{11}x^* = b_{02} + b_{12}x^*$. An alternative formulation of the estimation problem is in terms of a *test of constancy in the regression model* (BROWN et al. (1975)):

$$(3)\quad E[Y_t] = b_{0t} + b_{1t}x_t,\ t = 1, \ldots, T.$$

We test the null hypothesis $H_0: \underline{b}_t = \underline{b}$ for all t against the alternative hypothesis $H_1: \underline{b}_1 = \underline{b}$ for $t \le r^*$ and $\underline{b}_t \neq \underline{b}$ for $t > r^*$ for an index $r^* \in \{2, \ldots, T-1\}$, if $\underline{b}_t = (b_{0t}, b_{1t})$. The normal distribution with homogenous variance σ^2 is used as error distribution.

2.1 Sequential Regressions

Basic to many procedures in change-point regression is the calculation of sequences of regressions by adding or deleting one pair of observations (x_t, y_t). The sequence of regressions from the left is then given by the T-2 regressions

$$LR_{(r)}: y_t = b_0^{(r)} + b_1^{(r)}x_t + e_t,\ t = 1, \ldots, r,$$

with $x_{(r)} = (x_1, \ldots, x_r)$, $y_{(r)} = (y_1, \ldots, y_r)$

for $r = 3, \ldots, T$. The corresponding sequence of regressions from the right is given by the T-2 regressions

$$LR_{[r]}: y_t = b_0^{[r]} + b_1^{[r]}x_t + e_t,\ t = r+1, \ldots, T,$$

with $x_{[r]} = (x_{r+1}, \ldots, x_T)$, $y_{[r]} = (y_{r+1}, \ldots, y_T)$

for $r = 0, 1, \ldots, T-3$. For convenience we give the ANOVA table of both regressions $LR_{(r)}$ and $LR_{[r]}$ in Table 1. The regressions $LR_{(r)}$ and $LR_{[r]}$ complement each other with respect to the T pairs of data (x_t, y_t). Throughout the paper we shall use the (r) and [r] notation for statistics belonging to $LR_{(r)}$ and $LR_{[r]}$ and we abbreviate the estimates of the regression coefficients, error variances, and correlations by $b_0^{(r)}$, $b_1^{(r)}$, $s^2_{(r)}$, $R_{(r)}$ and $b_0^{[r]}$, $b_1^{[r]}$, $s^2_{[r]}$, $R_{[r]}$, respectively. For details of linear regression we recommend DRAPER and SMITH (1983), pp. 1–55.

III. Estimators of the break point r*

3.1 Error variance s^2

$$r^*: s^2_{(r^*)} = \min\,(s^2_{(r)},\ r = 3, \ldots, T).$$

The residual sum of squares $SS_{(r)}(Res)$ measures the goodness of fit and dividing by with the number of degrees of freedom it is standardized to $s^2_{(r)}$, an estimate of the variance about the regression, representing the error by which any observed value of Y can be predicted for a given value of x from the linear relationship. Hence, r* is based on the best prediction by linear regression.

3.2 Correlation: R^2

$$r^*: R^2_{(r^*)} = \max\,(R^2_{(r)},\ r = 3, \ldots, T).$$

The product correlation describes that part of the total variance about the mean of the y's which is explained by the regression.

$$R^2_{(r)} = SS_{(r)}(Res)/SS_{(r)}(Mean)$$
$$= 1 - (r\text{-}2)\,[s^2_{(r)}/SS_{(r)}(Mean)].$$

Notice the concordance of R^2 and s^2 as long as $SS_{(r)}(Mean)$ does not vary much with r. Both criteria, s^2 and R^2 need no assumption about the function f in (1) but they have the tendency to give very small estimates r*, in the extreme $r^* = 3$. Therefore, it is necessary for some applications to postulate a boundary condition like $r^* \geqq r_0$, to modify the criterion by looking for the largest local extremum within some range of r values, or to look for those r* where the criterion exceeds some boundary.

3.3 Residual Sum of Squares: S^2_r

We calculate for all reasonable T-4 segmentations of $\{1, \ldots, T\}$ the regressions $LR_{(r)}$ and $LR_{[r]}$, define

$$S^2_r = \begin{cases} SS_{(r)}(Res) + SS_{[r]}(Res),\ r = 3, \ldots, T\text{–}3 \\ SS_{(r)}(Res),\ r = T \end{cases}$$

and chose

$$r^*: S^2_{r^*} = \min\,\{S^2_r,\ r = 3, \ldots, T\text{–}3 \text{ or } r = T\}.$$

The case r = T was included for the option of a non-existing break point. No continuity condition in a point between x_{r^*} and x_{r^*+1} was assumed. SCHULZE (1984) used the S^2_r criterion – except the inclusion of r = T – in a two phase growth model. ESTERBY and ELSHARAWI (1981) showed in the more general case of polynomial regression that this least square estimate r* maximizes the relative marginal likelihood being also a relative conditional likelihood. The maximum likelihood estimator of the change-point of a two-segmented linear regression is a function of S^2_r, (HINKLEY (1969)).

3.4 Quandt's Likelihood Ratio Criterion: Q_r

QUANDT (1958) proposed a maximum likelihood estimate for the change-point of a two-segmented regression by

$$r^*: Q_{r^*} = \max\,\{Q_r,\ r = 3, \ldots, T\text{-}3)$$

with $Q_r = -T(\sqrt{2\pi} + 0.5) - r\log((r\text{-}2)s^2_{(r)}/r) - (T\text{-}r)\log((T\text{-}r\text{-}2)s^2_{[r]}/(T\text{-}r))$.

This criterion can become irregular at the boundaries where the estimates of the error variances are based on only few cases.

3.5 Recursive residuals: W_r, ω_r, L_r

Recursive residuals w_r were introduced by BROWN et al. (1975) to test for constancy of a multiple regression. The square of the r-th recursive residual w^2_r is the increment of the residual sum of squares if the r-th observation is added in the analysis. In the case of simple linear regression this means

$$SS_{(r)}(Res) = SS_{(r\text{-}1)}(Res) + w^2_r$$

or

$$w^2_r = (y_r - \hat{b}_0^{(r\text{-}1)} - \hat{b}_1^{(r\text{-}1)2})/(1 + B^{(r)})$$

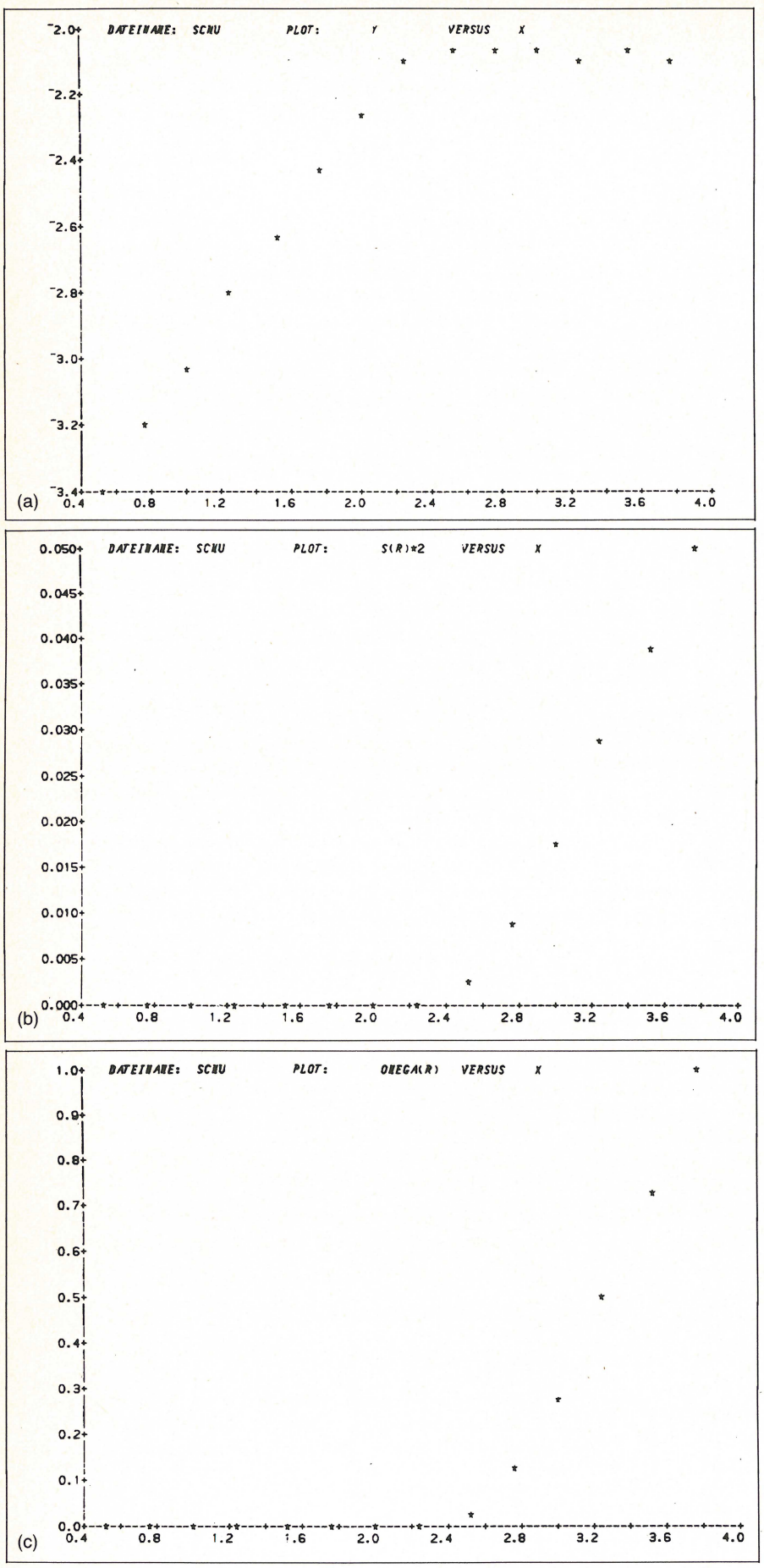
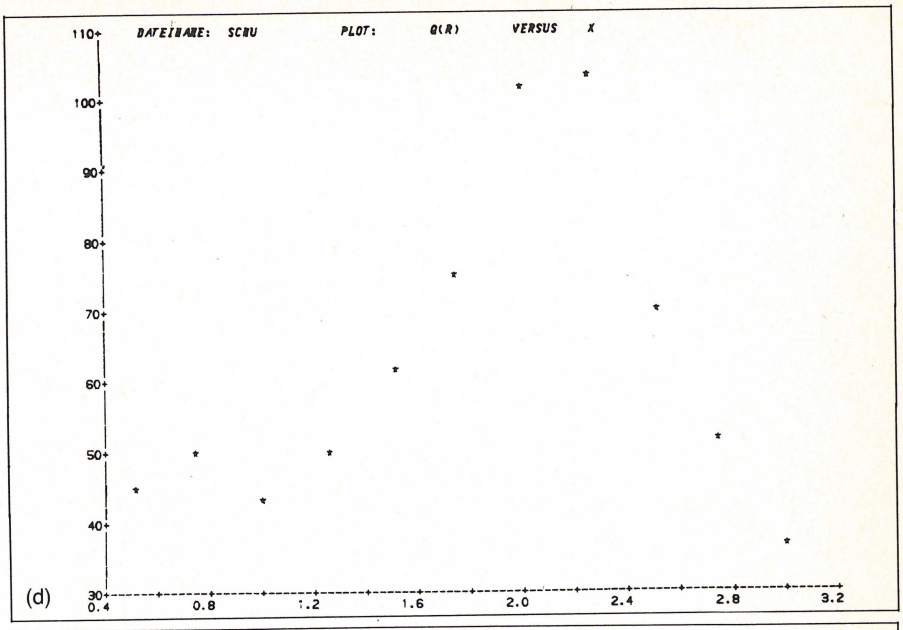
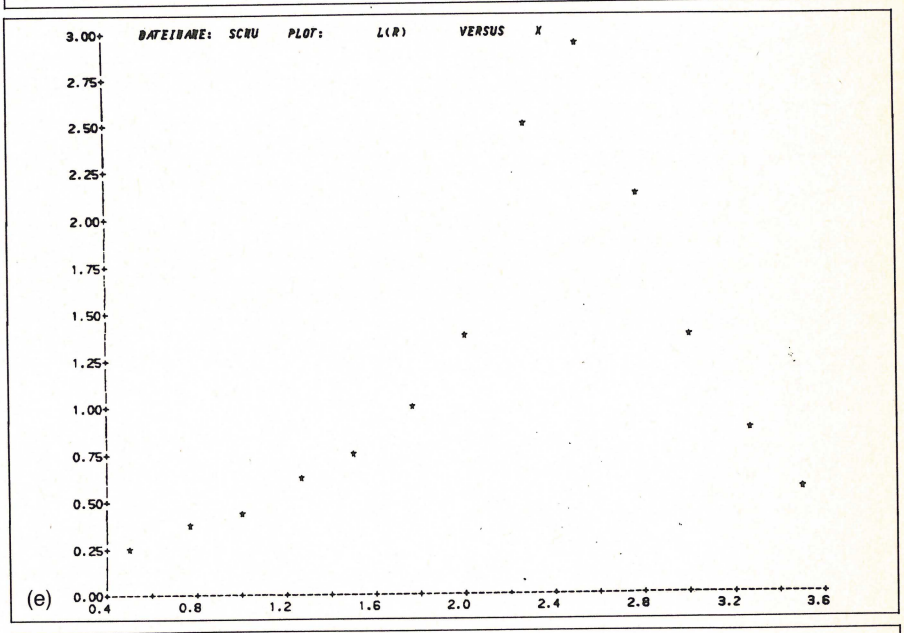


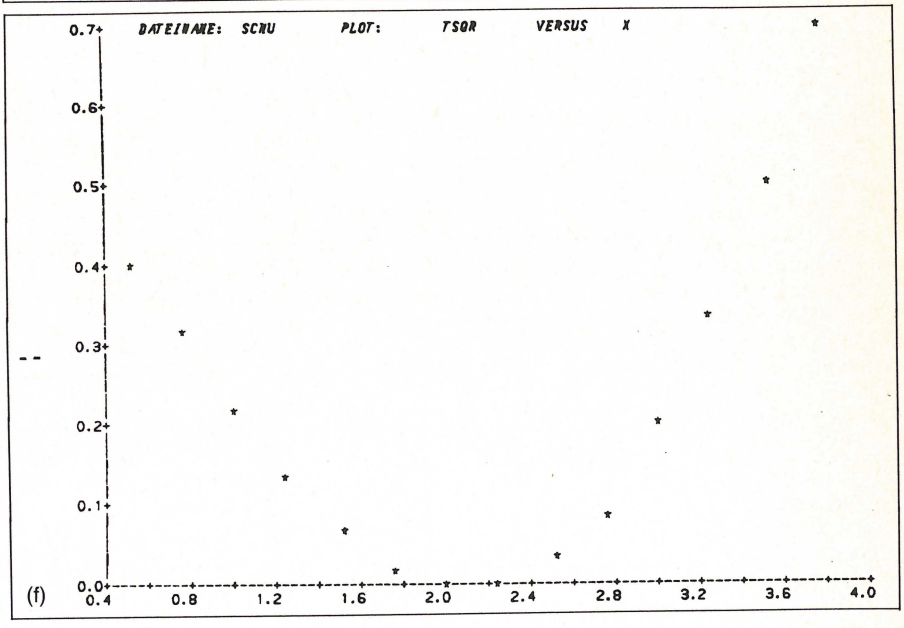
Fig. 1. Selected graphical results of CHREG if applied to the data of SCHULZE (1984) with plots of
(a) the dependent variable y versus the independent variable x
(b) the error variance versus x
(c) the cumulative sum of squares of recursive residuals versus x
(d) Quandt's maximum likelihood estimate versus x
(e) the likelihood ratio statistic for recursive residuals versus x
(f) the residual sum of squares versus x.


DATEINAME: SCHU PLOT: Q(R) VERSUS X
110
100
90
80
70
60
50
40
30
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2

(d)


DATEINAME: SCHU PLOT: L(R) VERSUS X
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6

(e)


DATEINAME: SCHU PLOT: TSQR VERSUS X
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0

(f)

with

$$B^{(r)} = (\Sigma_{i=1}^{r-1} x_i^2 - 2x_r \Sigma_{i=1}^{r-1} x_i + (r-1)x_r^2)/(r-1)\ SX_{(r-1)}$$

and

$$SX_{(r)} = \Sigma_{i=1}^{r}(x_i - \bar{x}_{(r)})^2$$

$r = 3, \ldots, T$.

Given the null hypothesis H_0 that (3) is constant for all t, w_3, w_4, . . . w_T are independent and identically normally distributed with expectation 0 and variance σ^2. Given the alternative H_1 this is true only for $t \leq r^*$. Two CUSUM-plots were suggested for a graphical assessment of r^*:

CRR = CUSUM of recursive residuals

$$W_r = \Sigma_{t=3}^{r} w_t/s_{(T)}$$

and

CSRR = CUSUM of squares of recursive residuals

$$\omega_r = \Sigma_{t=3}^{r} w_t^2/\Sigma_{t=3}^{T} w_t^2 = SS_{(r)}(Res)/SS_{(T)}(Res).$$

Increasing or decreasing straight line warning boundaries were obtained by BROWN et al. (1975) and critical regions of a formal test were given by DESHAYES and PICARD (1982). We used changes of the slopes in plots of W_r or ω_r versus x_r as indications for a change-point.

DESHAYES and PICARD (1982) proposed a likelihood ratio test of H_0 by the test-statistic

$$L_r = \frac{(\bar{w}_{3,r} - \bar{w}_{r+1,T})[(r-2)(1-(r-2)/(T-2))/(T-2)]^{1/2}}{[(SW_{3,r} + SW_{r+1,T})/(T-2)]^{1/2}}$$

for $r = 3, \ldots, T-1$, where

$$\bar{w}_{a,b} = \Sigma_{t=a}^{b} w_t/(b-a+1),$$

$$SW_{a,b} = \Sigma_{t=a}^{b}(w_t - \bar{w}_{a,b})^2.$$

We estimate

$$r^*: L_{r^*} = \max\{L_r, r = 3, \ldots, T-1\}.$$

3.6 Least Squares Residuals: z_r, Z_r

The residuals

$$e_t = y_t - \hat{b}_0^{(T)} - \hat{b}_1^{(T)} x_t$$

of the regression $LR_{(T)}$ based on all data were used by MCCABE and HARRISON (1980) to construct a test for constancy of a regression. They defined for $r = 3, \ldots, T$

CR = CUSUM of residuals

$$z_r = \Sigma_{t=1}^{r} e_t/(SS_{(T)}(Res))^{1/2}$$

and

CSR = CUSUM of squares of residuals

$$Z_r = \Sigma_{t=1}^{r} e_t^2/SS_{(T)}(Res) = \Sigma_{t=1}^{r} Res_{(r)}$$

and then looked for crossings with straight lines defined by a test of the null hypothesis H_0.

We considered the CUSUM methods as graphical estimation procedures. If C_r, $r = 3, \ldots, T$ denotes one of the CUSUMS from 3.5 or 3.6 it is convenient to use a standardized random walk presentation of the form $\{(k/T, C_{k+2}/T), k = 1, \ldots, T-2\}$ for a plot.

IV. Program CHREG

The sequential linear regressions $LR_{(r)}$ and $LR_{[r]}$ and the six criteria for the determination of a break point r^* reviewed in 3.1–3.6 were implemented in an interactive APL program CHREG. Because of the explorative nature of the statistical analysis considered here, the results of the different criteria were given explicitely in a numerical table as well as by printer plots. Table 2 lists the characteristics computed by the main program. Explicit estimates of r^* were given as described in 3.1–3.5. The range of search for r^* can be defined optionally in the dialog of CHREG.

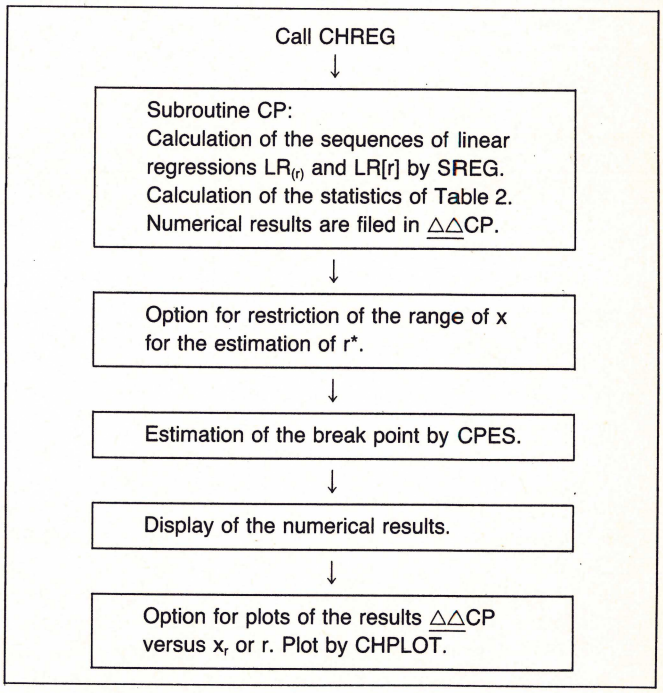


Table 1. Analysis of Variance (ANOVA) table for the sequences of linear regressions $LR_{(r)}$ and $LR_{[r]}$ if the total set of data is splitted into the first r (left) and the subsequent T–r (right) pairs of independent and dependent variables (x_r, y_r). We use the abbreviations $\bar{x}_{(r)} = (\Sigma_{i=1}^{r} x_i)/r$, $\bar{y}_{(r)} = \Sigma_{i=1}^{r} y_i)/r$, $\bar{x}_{[r]} = (\Sigma_{i=r+1}^{T} x_i)/T-r$, $\bar{y}_{[r)]} = (\Sigma_{i=r+1}^{T} y_i)/T-r$. $\hat{y}_i^{(r)}$ and $\hat{y}_i^{[r]}$ denote the fitted values in the respective regressions.

Source of Variation	d.f.	Sum of Squares	Mean Square
Left: $LR_{(r)}$			
Due to Regression	1	$SS_{(r)}(Reg) = \Sigma_{i=1}^{r}(\hat{y}_i^{(r)} - \bar{y}_{(r)})^2$	
About Regression	r–2	$SS_{(r)}(Res) = \Sigma_{i=1}^{r}(\hat{y}_i^{(r)} - y_i)^2$	$s^2_{(r)} = SS_{(r)}(Res)/(r-2)$
About Mean	r–1	$SS_{(r)}(Mean) = \Sigma_{i=1}^{r}(y_i - \bar{y}_{(r)})^2$	
Right: $LR_{[r]}$			
Due to Regression	1	$SS_{[r]}(Reg) = \Sigma_{i=r+1}^{T}(\hat{y}_i^{[r]} - \bar{y}_{[r]})^2$	
About Regression	T–r–2	$SS_{[r]}(Res) = \Sigma_{i=r+1}^{T}(\hat{y}_i^{[r]} - y_i)^2$	$s^2_{[r]} = SS_{[r]}(Res)/(T-r-2)$
About Mean	T–r–1	$SS_{[r]}(Mean) = \Sigma_{i=r+1}^{T}(y_i - \bar{y}_{[r]})^2$	

4.1 Flow Chart of CHREG

CHREG consists of a computation and of an evaluation part based on some few subroutines, see the flow chart below. CHREG requires the regression data (x_t, y_t) as a matrix with T rows and 2 columns.

4.2 Subroutines

Subroutine CP calculates the quantities listed in Table 2. The results are displayed with comments and stored as a global variable $\underline{\triangle\triangle}$ CP «name» for further processing. «name» is specified by the user. $\underline{\triangle\triangle}$CP is a matrix with T-2 rows corresponding to the range $(3, \ldots, T)$ and 21 columns. * was inserted if no value could be calculated for systematic reasons. Subroutine CPES determines the change-point estimates. The plots are generated interactively by the subroutine CHPLOT which grasps to $\underline{\triangle\triangle}$CP «name». Both, CP and CHPLOT, can be called autonomously. CP e.g. requires the data as a Tx2 matrix as right argument with the x_r monotone decreasing or increasing in the first column. The stepwise linear regressions $LR_{(r)}$ and $LR_{[r]}$ are calculated by the subroutine SRES in CP. SRES itself produces as result a Tx10 matrix with $\bar{x}_{(r)}$, $\bar{y}_{(r)}$, $SX_{(r)}$, $SY_{(r)}$, $SXY_{(r)}$, $\hat{b}_0^{(r)}$, $b_1^{(r)}$, $SS_{(r)}(Res)$, $R^2_{(r)}$, and $s^2_{(r)}$. SX, SY and SXY denote the respective sums of squares and products in the linear regression, see Table 1. The first two rows contain dummy results.

V. Example

CHREG was applied to the data of SCHULZE (1984). The numerical output is shown in Table 3. The error variance criterion located a break point at $x = 1.75$ whereas maximal correlation and minimal total residual sum of squares was at the subsequent $x = 2.0$. Quandt's maximum likelihood estimate was at the next larger point $x = 2.25$ and the largest change-point estimate was obtained by Deshayes's and Picard's likelihood ratio statistic at $x = 2.5$. A few plots are shown in Figure 1. The data themselves would suggest a break at $x = 2.25$ or 2.5. This was confirmed by the plot of SQRES(R). The recursive residuals suggested a break again between 2.25 and 2.5, whereas the cumulative sums of the residuals gave a point even larger than 2.5.

VI. Discussion

CHREG was designed as a computational statistical tool for biostatisticians forced to analyze time/dose-response data. A series of estimation procedures was implemented for an exploratory use. No universal or automatic program was intended which could be used by statistically untrained personal and which would give an unambiguous and definite answer to a given data set. CHREG, instead, gives a series of answers and it is the task of the biostatistician to arrive at a valid conclusion *with the results of CHREG* not disregarding the biological problem. We think the program can serve as an important help in this decision process and can be used as a module in an automatic advisory system (in modern terms: expert system) requiring the determination of break points.

VII. Hardware/Software Specifications and Availability

CHREG, the subroutines and some programs supporting in/output were implemented as an APL workspace CHREG with total size of about 100 kilobytes. The programs run on an IBM 4381 computer at the DKFZ in Heidelberg with VM system utilizing IBM APL Release 4. A source listing of the programs is available from the authors.

Table 2. Regression characters calculated by CHREG. The numbering by NO corresponds exactly to the options menue for plots of the characteristics versus x_r.

NO	Symbol	Regression characteristic
1.	x_r	independent variable
2.	y_r	dependent variable
3.	$b_0^{(r)}$	regression constant
4.	$b_1^{(r)}$	regression coefficient
5.	$SS_{(r)}(Res)$	sums of squares of residuals
6.	$R_{(r)}$	correlation coefficient
7.	$b_0^{[r]}$	regression constant backwards
8.	$b_1^{[r]}$	regression coefficient backwards
9.	$SS_{[r]}(Res)$	sums of squares of residuals backwards
10.	$R_{[r]}$	correlation coefficient backwards
11.	$s^2_{(r)}$	estimated error variance
12.	$s^2_{[r]}$	estimated error variance backwards
13.	w_r	recursive residuals
14.	W_r	cumulative sum of recursive residuals
15.	ω_r	cumulative sum of squares of recursive residuals
16.	Q_r	Quandt's maximum likelihood estimator
17.	L_r	likelihood ratio test statistic for recursive residuals
18.	z_r	cumulative sum of residuals
19.	Z_r	cumulative sum of squares of residuals
20.	$Res_{(r)}$	standardized squares of residuals
21.	S^2_r	$= SS_{(r)}(Res) + SS_{[r]}(Res)$

Table 3. Total numerical results of CHREG if applied to the data of SCHULZE (1984).

X	Y	BETA0(R)	BETA1(R)	SQRES(R)	R(R)*2	BETA0[R]
.5000	¯3.3903	¯3.7812	.7815	.00000	1.00000	¯3.2220
.7500	¯3.1991	¯3.7804	.7767	.00001	.99997	¯3.0900
1.0000	¯3.0345	¯3.7743	.7521	.00038	.99892	¯2.9182
1.2500	¯2.8134	¯3.7782	.7640	.00059	.99908	¯2.7355
1.5000	¯2.6409	¯3.7767	.7603	.00063	.99938	¯2.5158
1.7500	¯2.4476	¯3.7765	.7598	.00063	.99959	¯2.2891
2.0000	¯2.2740	¯3.7738	.7552	.00081	.99962	¯2.0838
2.2500	¯2.0972	¯3.7705	.7503	.00115	.99960	¯2.0503
2.5000	¯2.0826	¯3.7449	.7161	.02521	.99290	¯2.0242
2.7500	¯2.0747	¯3.7065	.6701	.08831	.97847	¯2.0184
3.0000	¯2.0770	¯3.6605	.6199	.19351	.95761	¯2.0251
3.2500	¯2.0883	¯3.6099	.5693	.33891	.93150	***
3.5000	¯2.0770	¯3.5601	.5233	.49921	.90567	***
3.7500	¯2.0972	¯3.5087	.4793	.69194	.87584	***

BETA1[R]	SQRES[R]	R[R]*2	S(R)*2	S[R]*2	W(R)	WC(R)
.3720	.40646	.7948	.00000	.03695	¯.0006	¯.0026
.3255	.31600	.7497	.00000	.03160	¯.0022	¯.0072
.2656	.21268	.6967	.00013	.02363	¯.0195	¯.0947
.2057	.13618	.6156	.00015	.01702	¯.0143	¯.0304
.1344	.06633	.5052	.00013	.00948	¯.0063	¯.0586
¯.0628	.02149	.3252	.00010	.00358	¯.0011	¯.0634
¯.0003	.00055	.0004	.00012	.00011	¯.0135	¯.1244
¯.0104	.00025	.3209	.00014	.00006	¯.0184	¯.2069
¯.0180	.00017	.5496	.00280	.00006	¯.1551	¯.9047
¯.0197	.00016	.4235	.00883	.00008	¯.2512	¯2.0346
¯.0178	.00016	.1949	.01759	.00016	¯.3243	¯3.4935
***	.00000	1.0000	.02824	***	¯.3813	¯5.2087
***	***	***	.03840	***	¯.4004	¯7.0096
***	***	***	.04942	***	¯.4390	¯8.9844

OMEGA(R)	Q(R)	L(R)	Z(R)	ZC(R)	RES(R)	TSQRES
.00000	45.0683	.2396	¯.7104	.1847	.0212	.40646
.00001	49.8987	.3612	¯.7704	.1883	.0036	.31601
.00055	42.6741	.4568	¯.7764	.1884	.0000	.21306
.00085	50.2385	.6135	¯.6608	.2017	.0134	.13677
.00091	61.3243	.7745	¯.4817	.2338	.0321	.06695
.00091	74.8607	1.0161	.2144	.3053	.0715	.02212
.00117	101.9082	1.3994	.1176	.4155	.1102	.00136
.00186	103.0921	2.4769	.5181	.5759	.1604	.00140
.03644	70.2997	2.9418	.7920	.6509	.0750	.02538
.12763	51.2205	2.1092	.9315	.6704	.0195	.08848
.27966	36.0287	1.3779	.9241	.6704	.0001	.19367
.48980	***	.8864	.7592	.6976	.0272	.33891
.72146	***	.5502	.4637	.7849	.0873	.49921
1.00000	***	***	.0000	1.0000	.2151	.69194

VIII. Acknowledgements

We are very much obliged to WERNER RITTGEN for giving us his powerful plot procedure at our disposal and to REGINA GRUNERT and RENATE RAUSCH for their help in preparing this manuscript.

Literature

BROWN, R. L., J. DURBIN, J. M. EVANS: Techniques for testing the constancy of regression relationships over time (with discussion). J. Roy. Statist. Soc. **B 37,** 149–195 (1975).

DESHAYES, J., D. PICARD: Testing for a change-point in statistical models. Report, Dept. Math., Univ. Paris-Sud, 91405 Orsay (1980).

DESHAYES, J., D. PICARD: Tests of disorder of regression: asymptotic comparison. Theory Prob. Appl. **27,** 100–115 (1982).

DRAPER, N. R., H. SMITH: Applied regression analysis. 2nd ed., New York: Wiley (1983).

ESTERBY, S. R., A. EL-SHAARAWI: Inference about the point of change in a regression model. Appl. Statist. **30,** 277–285 (1981).

HINKLEY, D. V.: Inference about the intersection in two-phase regression. Biometrika **56,** 495–504 (1969).

HINKLEY, D. V.: Inference about the change-point in a sequence of random variables. Biometrika **57,** 1–17 (1970).

HUDSON, D. J.: Fitting segmented curves whose join points have to be estimated. J. Amer. Statist. Assoc. **61,** 1097–1129 (1966).

MCCABE, B. P. M., M. J. HARRISON: Testing the constancy of regression relationships over time using least squares residuals. Appl. Statist. **29,** 142–148 (1980).

QUANDT, R. E.: The estimation of the parameters of a linear regression system obeying two separate regimes. J. Amer. Statist. Assoc. **53,** 873–880 (1958).

QUANDT, R. E.: Tests of the hypothesis that a linear regression system obeys two separate regimes. J. Amer. Statist. Assoc. **55,** 324–330 (1960).

SCHULZE, U.: A method of estimation of change points in multiphase growth models. Biom. J. **26,** 495–504 (1984).

Anschrift der Verfasser: Dr. Lutz Edler und Jutta Berger, Deutsches Krebsforschungszentrum Heidelberg, Institut für Dokumentation, Information und Statistik, Abteilung Biostatistik, Im Neuenheimer Feld 280, 6900 Heidelberg.

EDV in Medizin und Biologie **16** (4), 134–139, ISSN 0300-8282

Program design of two-sample tests for the analysis of right-censored data

H. Mollner, R. Haux and M. Schumacher

Summary

In this paper we want to point out how program development can be done for two-sample tests for the analysis of right-censored data. In a first step we derive a uniform description for all test statistics considered. In a second step we apply the technique of stepwise reduction of data structures.

Key words: *censored data, generalized linear rank tests, Renyi test, Kolmogorov-Smirnov-test, Cramér-von Mises test, program design, computational statistics.*

Zusammenfassung

In der Arbeit soll aufgezeigt werden, wie man Programme für Zweistichprobentests zur Analyse rechtszensierter Daten entwerfen kann. Zunächst leiten wir eine einheitliche Darstellung der untersuchten Teststatistiken her. Danach wenden wir die Technik der schrittweisen Reduktion von Datenstrukturen an.

Schlüsselwörter: *zensierte Daten, generalisierte lineare Rangtests, Renyi-Test, Kolmogorov-Smirnov-Test, Cramér-von-Mises-Test, Programmentwurf, rechnergestützte statistische Auswertungen.*

1. Introduction

Careful programming is a non-trivial subject in the field of computational statistics and we feel that Wirth's statement: ». . . the programmer's knowledge must not consist of a bag of tricks and trade secrets, but of a general intellectual ability to tackle problems systematically . . .« hits a very important point (WIRTH, 1976). For several classes of two-sample tests for the analysis of right-censored data we want to point out how program development can be done in order to pursue the above mentioned goal. Therefore we first derive a uniform description of these tests and show that they can be looked as one general test from which each single test can be regarded as a special case. The program design will be mainly based on applying the »stepwise reduction of data structures«. We show that this proceeding leads to a simple, intellectually manageable program which is versatile and which can support a user or a programmer with careful error diagnostics in order to improve the statistical quality of the results.

2. Basic definitions

2.1 *Statistical model*

In the sequel we consider the following situation: Let $X^0_{11},\ldots,X^0_{1N_1}$ and $X^0_{21},\ldots,X^0_{2N_2}$ be independent positive random variables representing the survival times or times to a certain event of N_1 individuals in group 1 and N_2 individuals in group 2, and let $N = N_1 + N_2$. We assume that the distribution functions F_k in the k-th group

$$F_k(t) = \Pr(X^0_{ki} \le t) ,$$

$k = 1, 2$, are absolutely continuous and strictly increasing. The corresponding survivor, hazard and cumulative hazard functions in the k-th group are denoted by S_k, λ_k and Λ_k, respectively.

Furthermore, the X^o_{ki} are censored on the right by independent positive random variables $C_{11},\ldots,C_{1N_1}$ and $C_{21},\ldots,C_{2N_2}$ with distribution functions G_1 and G_2.

These censoring variables C_{ki} are also assumed independent of the X^o_{ki}. Thus, in this model of random censorship one can only observe the random variables

$$X_{ki} = \min(X^o_{ki}, C_{ki})$$

and

$$\Delta_{ki} = 1_{\{X^0_{ki} \le C_{ki}\}} \qquad (k = 1, 2;\ i = 1,\ldots, N_k),$$

where Δ_{ki} indicates wheer X_{ki} is censored or not. Denote by $t_1 < t_2 < \ldots < t_j < \ldots < t_p$ the ordered, distinct uncensored survival times for the combined sample. Suppose that $D_k(t)$ deaths occur in the interval $[0, t]$ on group k,

$$\begin{aligned} D_k(t) &= \sum_{i=1}^{N_k} 1_{\{x_{ki} \le t \text{ and } \Delta_{ki} = 1\}} \\ &= D_{kj}\ (t_j \le t < t_{j+1}) , \end{aligned}$$

and $d_{kj} = D_{kj} - D_{kj-1}$ deaths at time t_j in the k-th group. The number of patients at risk at time $t-0$ in group k is given by $N_k(t)$, i.e.

$$\begin{aligned} N_k(t) &= \sum_{i=1}^{N_k} 1_{\{x_{ki} \ge t\}} \\ &= N_{kj}\ (t_j \le t < t_{j+1}). \end{aligned}$$

The cumulative hazard function $\Delta_k(t)$ can be estimated by NELSON (1969)

$$\begin{aligned} \hat\Lambda_k(t) &= \sum_{i:t_i \le t} d_{ki}/N_{ki} \\ &= \sum_{i:t_i \le t} \hat\lambda_{ki} \\ &= \hat\Lambda_{kj}\ (t_j \le t < t_{j+1}) \end{aligned} \qquad (2.1.1)$$

and the survivor function $S_k(t)$ by KAPLAN & MEIER (1958)

$$\begin{aligned} \hat S_k(t) &= \prod_{i:t_i \le t} (1 - d_{ki}/N_{ki}) \\ &= \prod_{i:t_i \le t} (1 - \hat\lambda_{ki}) \\ &= \hat S_{kj}\ (t_j \le t < t_{j+1}) \end{aligned} \qquad (2.1.2)$$

where $\hat\lambda_{ki} = d_{ki}/N_{ki}$ is an estimate of $\lambda_k(t)$ at time t_i. Our interested centres around whether the survival distributions in the two groups are equal or not, i. e. the test problem is given by

$$H_0 : S_1(t) = S_2(t) \text{ for all } t > 0$$

vs.

$$H_1 : S_1(t) \neq S_2(t) \text{ for at least one } t > 0 .$$

2.2 *Stepwise reduction of data structures*

The stepwise reduction of data structures is a programming technique which can be especially used for the development of programs for statistical data analysis. A short description of this technique can be found in HAUX (1982) and a more detailed one in HAUX (1984). Briefly, the technique can be characterized as follows:

(1) We first try to find all possible (and meaningful) input and output data structure groups.
(2) Then we construct the so-called reduction path(s) through which each output data structure group can be arrived at from every input data structure group.

By reduction we mean: there exists a definite rule that maps each instance of the one data structure group to an instance of another. Such a mapping rule can be given by a (sub-) program.

A data structure group is defined to consist of a sequence of data structures such as sequential files (data matrices), relations, arrays, records or scalars (for the term relation see CODD (1979), the other terms are used according to WIRTH, 1976). An input data structure group contains the data to be analyzed as well as parameters for the specification of the data processing, an output data structure group contains the results, a status report (error messages), etc.

The properties of this technique and its application to statistical analysis systems (as well as some simple examples) can be found in HAUX (1982, 1984) and will not be discussed here. We only want to mention that the stepwise reduction can be applied, if the programs are written in application oriented languages such as PASCAL, PL/I or FORTRAN. In this context each reduction step can be regarded as a subprogram whose parameter list contains the input and output data structure groups.

Note that in this paper we use the term »program« synonymously to the term »method« (of the methodbase of a statistical analysis system) in HAUX (1982, 1984).

3. Tests for the analysis of right-censored data

3.1 *Generalized linear rank tests and so-called »Renyi-type« statistics*

AALEN (1978) showed that the test statistics of the generalized linear rank tests can be written as

$$Q_J(\tau) = \int_0^\tau J(u)\ \left(d\hat\Lambda_2(u) - d\hat\Lambda_1(u)\right) \qquad (3.1.1)$$

where $J(t)$ denotes a positive weight function and τ the greatest uncensored observation at time t_p. A consistent estimator of the variance of $Q_J(\tau)$ is given by GILL (1980)

$$\widehat{\mathrm{var}}(Q_J(\tau)) = \sum_{k=1}^{2} \int_0^\tau (J(u))^2/N_k(u)\ d\hat\Lambda_k(u) \qquad (3.1.2)$$

Since the integrals in (3.1.1) and (3.1.2) reduce to finite sums, we get

$$Q_J(\tau) = Q_{J_p} = \sum_{j=1}^{p} J_j(\hat\lambda_{2j} - \hat\lambda_{1j}) \qquad (3.1.3)$$

and

$$\widehat{\mathrm{var}}(Q_J(\tau)) = \widehat{\mathrm{var}}(Q_{J_p}) = \sum_{j=1}^{p} (J_j)^2\ (\hat\lambda_{2j}/N_{2j} + \hat\lambda_{2j}/N_{1j}). \qquad (3.1.4)$$

The standardized test statistic of the generalized linear rank test has under H_0 asymptotically ($N_1, N_2 \to \infty$) a standard normal distribution. If we choose as weight function

$$J_j = N_{ij}\ N_{ij}/(N_{1j} + N_{2j})$$

we obtain the well-known logrank test (PETO & PETO, 1972)

$$J_j = N_{1j}\, N_{2j}$$

we get the generalized Wilcoxon test proposed by GEHAN (1965). HARRINGTON and FLEMING (1982) proposed a class of weight functions

$$J_j = (\hat{S}_{\cdot j})^p\, N_{1j}\, N_{2j}/(N_{1j} + N_{2j})$$

with $p \geq 0$, where $\hat{S}_{\cdot j}$ denotes the Kaplan-Meier estimator of the combined sample. For $p = 0$ we obtain again the logkrank test and for $p = 1$ we get a version of the generalized Wilcoxon test as proposed by PRENTICE (1978).

The so-called »Renyi-type« statistics use the maximum of the distance between the two empirical processes

$$\int_0^t J(u)\, d\hat{\Lambda}_1(u) \text{ and } \int_0^t J(u)\, d\hat{\Lambda}_2(u)$$

as measure of the difference between the survival distributions. These test statistics are based on the generalized linear rank tests and are defined by

$$Q_{GJ} = \max_{0 \leq j \leq p} \left[\frac{Q_{J_j}}{(\widehat{\mathrm{var}}(Q_{J_j}))^{1/2}} \right] \tag{3.1.5}$$

and

$$Q^0_{GJ} = \max_{0 \leq j \leq p} \left[\frac{Q_{J_j}}{1 + \widehat{\mathrm{var}}(Q_{J_j})} \right] \tag{3.1.6}$$

When we use the above mentioned weight functions we obtain the maxima of logrank, Gehan, Harrington-Fleming and Prentice tests, each of them successively calculated in time.

3.2 *Test statistics of Kolmogorov-Smirnov and Cramér-von Mises type*

For describing the difference of the two survival distributions we can consider the function

$$\xi(t) = \Lambda_2(t) - \Lambda_1(t), \tag{3.2.1}$$

i.e. the difference of the cumulative hazard functions. In particular it is equal to zero if the two survival distributions are equal. Thus the test problem can also be written as

H_0: $\xi(t) = 0$ for all $t > 0$

vs.

H_1: $\xi(t) \neq 0$ for at least one $t > 0$.

The difference of the cumulative hazard functions can be estimated by

$$\hat{\xi}(t) = \hat{\Lambda}_2(t) - \hat{\Lambda}_1(t) \text{ or}$$
$$\hat{\xi}_j = \Lambda_{2j} - \Lambda_{1j},\ t_j \leq t < t_{j+1} \tag{3.2.2}$$

and a measure for the deviation of $\hat{\xi}(t)$ from the zero-line can be used as a suitable test statistic.

An application of the theorems of BRESLOW & CROWLEY (1974) yields that the asymptotic variance of $\hat{\xi}(t)$ is approximately equal to

$$\hat{A}_{\cdot j} = \frac{N}{N_1}\hat{A}_{1j} + \frac{N}{N_2}\hat{A}_{2j},\ t_j \leq t < t_{j+1},$$

where $\hat{A}_{kj}$ is an estimate of the so-called censoring integral in the k-th group,

$$A_k(t) = \int_0^t \frac{dF_k(u)}{(S_k(u))^2(1-G_k(u))}$$

The estimate

$$\hat{A}_k(t) = N_k \sum_{i:t_i \leq t} \frac{d_{ki}}{(N_{ki} - d_{ki})N_{ki}}$$
$$= \hat{A}_{kj},\ t_j \leq t < t_{j+1}$$

was proposed by AALEN (1976) and HALL & WELLNER (1980).

This leads to the following test statistics (SCHUMACHER, 1984)

$$Q_{KS\xi} = \sqrt{N} \max_{0 \leq j \leq p^*} \left[\frac{\hat{\xi}_j}{(\hat{A}_{\cdot j})^{1/2}} \right] \tag{3.2.3}$$

and

$$Q^0_{KS\xi} = \sqrt{N} \max_{0 \leq j \leq p^*} \left[\frac{\hat{\xi}_j}{(1 + \hat{A}_{\cdot j})} \right] \tag{3.2.4}$$

of Kolmogorov-Smirnov type and

$$Q_{CM\xi} = N \sum_{j=1}^{p^*} (\hat{\xi}_j/\hat{A}_{\cdot p^*})^2 \quad (\hat{A}_{\cdot j} - \hat{A}_{\cdot j-1}) \tag{3.2.5}$$

and

$$Q^0_{CM\xi} = N \sum_{j=1}^{p^*} (\hat{\xi}_j/(1+\hat{A}_{\cdot j}))^2 \quad (\hat{H}_{\cdot j} - \hat{H}_{\cdot j-1}) \tag{3.2.6}$$

of Cramér-von Mises-type where $\hat{H}_{\cdot j} = \hat{A}_{\cdot j}/(1+\hat{A}_{\cdot j})$. The number of events at time $\tau^* = \sup\{t: \min(N_1(t), N_2(t)) > 1\}$ is denoted by p^*.

Another function that describes the difference between the two survival distributions is the »log-effect function« $\beta(t) = \log(\Lambda_2(t)) - \log(\Lambda_1(t))$.

Using similar arguments as above one obtains that the asymptotic variance of $\hat{\beta}(t)$ is approximately equal to $\hat{A}_{\cdot j}/(\hat{A}_{\cdot j})^2$, $t_j \leq t < t_{j+1}$, where $\hat{A}_{\cdot j}$ denotes the estimator of the cumulative hazard function in the combined sample. This leads to the test statistics (SCHUMACHER, 1984)

$$Q_{KS\beta} = \sqrt{N} \max_{q \leq j \leq p^*} \left[\frac{\hat{A}_{\cdot j}\, \hat{\beta}_j}{(\hat{A}_{\cdot p^*})^{1/2}} \right] \tag{3.2.7}$$

and

$$Q^0_{KS\beta} = \sqrt{N} \max_{q \leq j \leq p^*} \left[\frac{\hat{\lambda}_{\cdot j}\, \hat{\beta}_j}{(1 + \hat{A}_{\cdot j})} \right] \tag{3.2.8}$$

of Kolmogorov-Smirnov type and

$$Q_{CM\beta} = N \sum_{j=q}^{p^*} (\hat{A}_{\cdot j}\, \hat{\beta}_j/\hat{A}_{\cdot p^*})^2 (\hat{A}_{\cdot j} - \hat{A}_{\cdot j-1}) \tag{3.2.9}$$

and

$$Q^0_{CM\beta} = N \sum_{j=q}^{p^*} (\hat{A}_{\cdot j}\, \hat{\beta}_j/(1+\hat{A}_{\cdot j}))^2 (\hat{H}_{\cdot j} - \hat{H}_{\cdot j-1}) \tag{3.2.10}$$

of Cramér-von Mises type where $\delta = \min\{t > 0: \hat{\Lambda}_1(t) > 0 \text{ and } \Lambda_2(t) > 0\}$ correspond to t_q.

A Cramér-von Mises type stastistic based on the difference $\eta(t) = S_2(t) - S_1(t)$ can be constructed in the same way. Defining

$\hat{\eta}(t) = \hat{S}_2(t) - \hat{S}_1(t)$, $\hat{\eta}_j = \hat{S}_{2j} - \hat{S}_{1j}$, $t_j \leq t < t_{j+1}$, we obtain the test statistic

$$Q^0_{CM\eta} = N \sum_{j=1}^{p^*} (\hat{S}_{\cdot j}\, \hat{\eta}_j/(1+\hat{A}_{\cdot j}))^2 (\hat{H}_{\cdot j} - \hat{H}_{\cdot j-1}) \tag{3.2.11}$$

where $\hat{S}_{\cdot j}$ denotes the Kaplan-Meier estimate of the survivor function in the combined sample (KOZIOL & YUH, 1982).

Details on the derivations and on the distributions of the test statistics under the null hypothesis can be found in SCHUMACHER (1984).

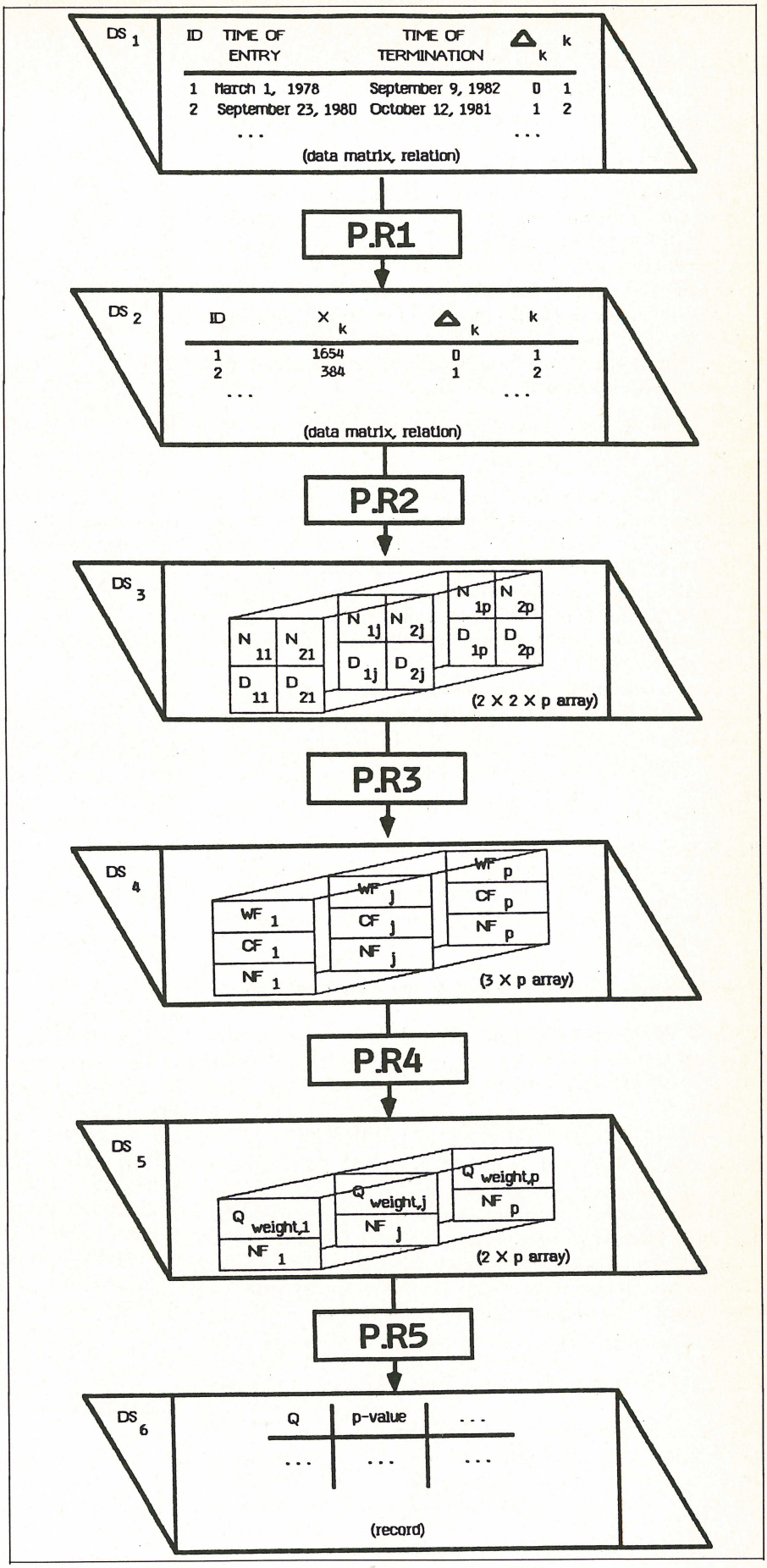


Figure 1. Reduction diagram for the generalized test.

4. Uniform description of the tests

The generalized linear rank tests as well as the Renyi-type tests can be expressed as

$$Q_1 = f\big((J_1,\hat{\lambda}_{21}-\hat{\lambda}_{11},\widehat{var}(Q_{J1})),\ldots,(J_p,\hat{\lambda}_{2p}-\hat{\lambda}_{1p},\widehat{var}(Q_{J_p}))\big)$$
$$= f\big((Q_{J1},\widehat{var}(Q_{J_1})),\ldots,(Q_{J_1},\widehat{var}\,(Q_{J_p})\big), \quad (4.1)$$

which means that the test statistic is a function of the weight function J_j, of the difference of the empirical hazard functions $\hat{\lambda}_{2j}-\hat{\lambda}_{1j}$ and of the variance $\widehat{var}(Q_{Jj})$, or on $Q_{Jj}=J_j(\hat{\lambda}_{2j}-\hat{\lambda}_{1j})$ and $\widehat{var}\,(Q_{Jj})$, $j=1,\ldots,p$, respectively.

For the Kolmogorov-Smirnov and Cramér-von Mises tests we can find an analoguous representation if we denote as $\hat{\psi}_j$ either the difference of the empirical cumulative hazard function $\hat{\xi}_{sj}$, the empirical log-effect function $\hat{\beta}_j$, or the difference of the Kaplan-Meier estimates. We obtain

$$Q_2 = f\,((\phi_1,\hat{\psi}_1,\hat{A}_{\cdot 1}),\ldots,(\phi_p,\hat{\psi}_p,\hat{A}_{\cdot p}))$$
$$= f\,((Q_{\phi_1},\hat{A}_{\cdot 1}),\ldots,(Q_{\phi_p},\hat{A}_p)) \quad (4.2)$$

The »weight« ϕ_i is equal to 1 for test statistics which are based on the difference of the empirical cumulative hazard functions, $\hat{\xi}(t)$, in (3.2.3) to (3.2.6). For the test statistics (3.2.7) to (3.2.10) the ϕ_j are equal to the Nelson estimate $\hat{\Lambda}_{\cdot j}$ of the cumulative hazard function in the combined sample and for the test statistic of the Cramér-von Mises type, $Q_{CM\eta}$, in (3.2.11) we have $\phi_j=\hat{S}_{\cdot j}$ which corresponds to the Kaplan-Meier estimator of the combined sample.

Now it is obvious to choose a uniform description of the test statistics which contains (4.1) as well as (4.2). Therefore we take

$$Q = f\,((WF_1,CF_1,NF_1),\ldots,(WF_p,CF_p,NF_p))$$
$$= f\,((Q_{weight,1},NF_1),\ldots,(Q_{weight,p},NF)_p) \quad (4.3)$$

and denote a weight function (J_j or ϕ_j) as WF_j, the difference of the functions characterizing the two survival distributions ($\hat{\lambda}_{2j}-\hat{\lambda}_{1j}$ or ψ_j) as CF_j, and a normalizing function ($\widehat{var}\,(Q_{J_j})$ or $\hat{A}_{\cdot j}$) as NF_j (MOLLNER, 1983). We can denote Q as our general test statistic.

5. Application of the technique of stepwise reduction

Because of the uniform description of the test statistics as shown in the last section, we only need to apply the stepwise reduction of data structures to the one general test.

5.1 Possible input and output data structure groups

At first let us define all possible and meaningful input and output data structure groups for the general test. For the sake of simplicity we will not mention data structures, containing e.g. indications which special tests should be computed.

As first input data structure (»group« will be omitted if there is only one structure in the group) for the test we can specify a case-oriented one which contains for every case or experimental unit, respectively, the identification, the time of entry into the study, the time of termination, an indicator for censoring (Δ_k) and the treatment (k). Let us denote this data structure as DS_1.

A second possible input data structure, DS_2 say, is a structure in which the observed time-on-study (X_k), i.e. the difference between entry and termination time, is given instead of the dates of entry and termination. DS_2 is also case-oriented.

Both data structures can be represented in relations or in sequential files (data matrices). These two input data structures are e.g. supported in BMDP1L and BMDP2L (DIXON, 1981).

The next possible input data structure, denoted as DS_3, is a 3-dimensional array, containing the functions N_{kj} and D_{kj}, $k=1,2$, $j=1,\ldots,p$. The values of DS_3 cannot be assigned to one case; DS_3 is a not case-oriented structure.

DS_1, DS_2 and DS_3 are possible and meaningful input data structures for a user. However, we can define additional input data structures. The following structures are all not case-oriented. The next two will not be of importance for a user. But they can help a programmer to modularize a program properly in order to keep it clearly written and intellectually manageable (WIRTH, 1974).

The first one of these two, DS_4 say, consists of a 2-dimensional $3\times p$-array, containing for each time t_j, $j=1,\ldots,p$, the values of the weight-function WF_j, of the characteristic function CF_j and of the normalizing function NF_j. The next one, DS_5 say, is a $2\times p$-array containing the unnormed $Q_{weight,\,j}$ and the values of the normalizing function.

As a meaningful output data structure we can use a record which contains the value of the test statistic, the p-value, a status report (e.g. error messages), etc.

To our opinion DS_1 to DS_5 are the possible and meaningful input data structures and DS_6 is the possible output data structure of the general test.

5.2 Reduction path

We now have to find the reduction path(s) through which the output data structure can be reached from every input data structure. As in section 5.1 we omit the term »group« because, in each data structure group, there is only one data structure. According to HAUX (1984) let us first find out which data structure can be reduced to which other data structure. It can be easily seen that for $1\le i<j\le 6$ each data structure DS_i can be reduced to each DS_j, because for these pairs of data structures we know definite rules to get from DS_i to DS_j. Each of these rules can be implemented as a program.

Let us now drop all superfluous »transitive« rules, i.e. all rules that can be expressed by other ones. E.g. DS_3 can be reached from DS_1 by reducing first DS_1 to DS_2 and then DS_2 to DS_3. Therefore we need no »direct« reduction rule (DS_1, DS_3) to get from DS_1 to DS_3. After all superfluous rules have been removed the following rules remain: DS_1 to DS_2, DS_2 to DS_3, DS_3 to DS_4, DS_4 to DS_5 and DS_5 to DS_6. The (sub-)programs for implementing these rules will be denoted as P.R1, P.R2, P.R3, P.R4 and P.R5, where »P« stands for program and »R« for the degree of reduction (see figure 1).

P.R1 reduces DS_1 to DS_2 and then calls up P.R2, P.R2 reduces DS_2 to DS_3 and then calls up P.R3 etc. A user who has as input data structure DS_1 has to call up P.R1 to obtain the results; a user who has the data for the analysis already in the more reduced from of DS_2 only has to call up P.R2 etc.

By these five programs we can get from each input data structure DS_i, $i=1,\ldots,5$ to the output data structure DS_6 and we can now obtain, for data at any of the five given degrees of reduction, the desired results.

6. Discussion

By deriving a general test statistic we can limit the effort in programming. We now only have to program one general test instead of several special ones.

Because all (special) test statistics are based on $\hat{\Lambda}_k$ or $\hat{S}_k$ (2.1.1, 2.1.2) we always have to sort the values of the observed

survival times. Thus, the time complexity is for all tests at least as large as the one for sorting algorithms which belong to the polynomial class $O((N)^2)$. When we use the uniform description for computing the test statistics we only have besides sorting, appropriately to build sums or to find maxima. The time complexity therefore is less than or equal to $O((N)^2)$. Thus the time complexity of the program for the proposed general test belongs (at least) to the same polynomial class of each program for a special test.

By applying the stepwise reduction of data structures the procedure of obtaining the test statistic from a given input data structure becomes more intelligible and we can modularize the test in order to get a clearly written, intellectual manageable program. In addition, a number of possible data structures is now available for a user.

Within each of these modules other modularization techniques, like Wirth's stepwise refinement can (and should) be used, too. For each module, e.g. implemented as subprogram, we can specify appropriate semantic integrity constraints. Proceeding in such a way, we are able to obtain a substantial improvement of error checking (HAUX, 1983) and we therefore get a better support for a user. If we look at the modularized program not as a single object but as an element of the methodbase of a statistical analysis system we also get a more clearly structured methodbase (HAUX, 1984).

Finally let us note that the program design of the two-sample tests for the analysis of right-censored data, as described here, was of course somewhat simplified. So it was possible to focus on some principle ideas how careful programming can be tackled in the field of computational statistics.

Literature

O. O. AALEN (1976): Nonparametric inference in connection with multiple decrement models. Scand. Journ. Statist. **3,** 15–27.

O. O. AALEN (1978): Nonparametric inference for a family of counting processes. Ann. Statist. **6,** 701–726.

N. BRESLOW and J. CROWLEY (1974): A large sample study of the life table and product limit estimates under random censorship. Ann. Statist. **2,** 437–453.

E. F. CODD (1979): Extending the database relational model to capture more meaning. ACM Trans. Database Syst. **4,** 397–434.

W. J. DIXON (1981): BMDP Statistical Software 1981. Univ. of California Press, Berkeley.

E. GEHAN (1965): A generalized Wilcoxon test for comparing arbitrarily single censored samples. Biometrika **52,** 203–223.

R. D. GILL (1980): Censoring and stochastic integrals. Math. Centre Tracts **124,** Mathematisch Centrum Amsterdam.

W. J. HALL and J. A. WELLNER (1980): Confidence bands for a survival curve from censored data. Biometrika **67,** 133–143.

D. P. HARRINGTON and T. R. FLEMING (1982): A class of rank test procedures for censored survival data. Biometrika **69,** 553–566.

R. HAUX (1982): A programming technique for software in statistical analysis. In: H. CAUSSINUS, P. ETTINGER and P. TAMASSONE (Eds.), Compstat **82,** 266–271, Physika, Vienna.

R. HAUX (1983): How to detect and prevent errors in computer-supported statistical analysis: an example. Meth. Inform. Med. **22,** 87–92.

R. HAUX (1984): Statistical analysis system – construction and aspects of method design (part 2), Stat. Software Newsl. **10,** 14–27.

E. L. KAPLAN and P. MEIER (1958): Nonparametric estimation from incomplete observations. J. Am. Statist. Assoc. **53,** 457–481.

J. A. KOZIOL and Y. S. YUH (1982): Omnibus two-sample test procedures with randomly censored data. Biom. J. **24,** 743–750.

H. MOLLNER (1983): Vergleich von Zweistichprobentests für zensierte Daten, Diploma-thesis. Medical Informatics, Univ. of Heidelberg, FRG.

W. NELSON (1969): Hazard plotting for incomplete failure data. J. of Qual. Techn. **1,** 27–52.

R. PETO and J. PETO (1972): Asymptomatically efficient rank invariant test procedures (with discussion). J. Roy. Statist. Soc. A **135,** 185–206.

R. L. PRENTICE (1978): Linear rank tests with right censored data. Biometrika **65,** 167–179.

M. SCHUMACHER (1984): Two-sample tests of Cramér-von Mises- and Kolmogorov-Smirnov-type for randomly censored data. Int. Statist. Rev. **52,** 263–281.

N. WIRTH (1974): On the composition of well-structured programs. ACM Comp. Surveys **6,** 247–259.

N. WIRTH (1976): Algorithms + Data Structure = Programs. Prentice Hall, Englewood Cliffs, N.J.

Authors' addresses: Dipl.-Inform. Med. Hans Mollner. Dr. Reinhold Haux, Abteilung Medizinische Statistik und Dokumentation der RWTH Aachen, Pauwelsstraße, 5100 Aachen. Prof. Dr. Martin Schumacher, Fachbereich Statistik der Universität Dortmund, Postfach 50 05 00, 4600 Dortmund.

PERSÖNLICHES

Heinz Fink 65 Jahre

Am 7. Februar 1986 feiert Herr Professor Dr. med. Heinz FINK seinen 65. Geburtstag. Die Schar der Gratulanten, in die ich mich mit diesem Beitrag einreihen möchte, wird zwar nicht unübersehbar, aber doch sehr groß sein.

Seit 1977 hat er als Mitherausgeber der „EDV in Medizin und Biologie" diese Zeitschrift immer wieder entscheidend mitgeprägt. Auch dafür sei ihm heute sehr herzlich gedankt.

Sein Lebenslauf erscheint im nachhinein so normal, wie ein Lebenslauf bei den heute 60- bis 70jährigen nur normal sein kann. Dennoch wissen seine Freunde und Kollegen, daß nicht immer alles so glatt lief, daß er es sich und seinen Mitarbeitern in seiner Gradlinigkeit auch selbst nicht immer leicht machte. Wenn man aber bedenkt, in wie vielen Gremien und an wie vielen Stellen sein Rat und seine Mitarbeit gesucht wurde, wie aktiv er in wissenschaftlichen Gesellschaften war, dann wird deutlich, wie engagiert und produktiv er war. Hier fällt es schwer, alle Kontakte und Aktivitäten aufzuzählen.

Zwei Aktivitäten müssen aber angesprochen werden, da Professor FINK hier zu den „Pionieren" zu zählen ist. Zunächst ist es die Biometrie, mit der er sich schon 1948 bei seiner Promotion beschäftigte und die er dann nie mehr aus den Augen gelassen hat, gleichgültig, ob er in der Klinik oder in der Industrie tätig war. So war es nicht verwunderlich, daß er 1977–1979 zum Präsidenten der Deutschen Region der Biometrischen Gesellschaft gewählt wurde und von 1980–1983 die deutschen Mitglieder im Council dieser internationalen Gesellschaft vertreten durfte.

Einen zweiten Schwerpunkt könnte man in der Dokumentation sehen, wenn man die „Klinischen Prüfungen" noch bei der Biometrie subsummiert. Auch hier hat er von seiner Dissertation an die Entwicklung mitgetragen von den ersten Ansätzen der Lochkartenverfahren über die vielen Zwischenvarianten bis hin zu den heutigen Datenbankanwendungen.

So überrascht auch nicht seine Berufung 1972 in den Beirat des „Deutschen Instituts für Medizinische Dokumentation und Information (DIMDI)".

Wenn Herr Professor FINK auch mit der Erreichung der „Altersgrenze" aus dem aktiven Berufsleben ausscheidet, so wünsche ich ihm noch viel Zeit für die Beschäftigung mit seinem Hobby der Archäologie, und uns allen wünsche ich, daß wir noch recht oft mit ihm diskutieren können.

HANS GEIDEL

BUCHBESPRECHUNGEN

KÜFFNER, H. und WITTENBERG, R.
Datenanalyse für statistische Auswertungen – Eine Einführung in SPSS, BMDP und SAS
1985, 289 S., DM 36,–
G. Fischer Verlag, Stuttgart – New York

Als Vorstufe von Expertensystemen verwenden wir zur Zeit Programmpakete wie SPSS, BMDP, SAS und andere zur statistischen Datenanalyse. In der vorliegenden Einführung werden nach allgemeinen Anmerkungen zur empirischen Datenanalyse und einer Zusammenstellung der Voraussetzungen, Möglichkeiten und Grenzen der ausgewählten statistischen Auswertungsverfahren die drei Programmpakete (SPSS, BMDP und SAS) an einem konkreten, überschaubaren Datensatz (80 Datensätze) für die ausgewählten statistischen Methoden dargestellt. Viel Beachtung wird dabei den oft nicht leicht zu lesenden Computerausdrücken gewidmet, und dabei wird der Zusammenhang der ausgedruckten Zahlen mit den entsprechenden statistischen Parametern hergestellt. Hinweise auf das Datenbanksystem SIR und graphische Darstellungsmöglichkeiten runden die Darstellung ab.

Wenn auch schon spezielle Bedienungsanleitungen für die einzelnen Programmpakete vorliegen, so kann diese drei Pakete umfassende Darstellung vor allem für Benutzer größerer Rechenzentren nützlich sein, bei denen normalerweise nebeneinander mehrere Pakete verfügbar sind. Ge.

UNKELBACH, H. D. und WOLF, T.
Qualitative Dosis-Wirkungs-Analysen
1985, 138 S., DM 44,–
G. Fischer Verlag, Stuttgart – New York

Dieser zweite Band der Reihe „Biometrie" unterstreicht die Zweckmöglichkeit der Konzeption dieser Reihe. In einer Einzeldarstellung läßt sich ein Thema viel klarer und übersichtlicher behandeln, als es in einem umfassenden Lehrbuch möglich wäre.

Die vorliegende Darstellung der qualitativen Dosis-Wirkungs-Analyse behandelt zunächst die Analyse für eine Wirksubstanz, dann den Vergleich von Dosis-Wirkungs-Kurven sowie die Kombination von Substanzen. Im Anhang werden mathematische Formeln auch als Hilfsmittel zur Erstellung von Programmen zusammengestellt.

Diese klare, didaktisch gut aufgebaute Darstellung kann allen Interessenten wärmstens empfohlen werden und mag als Muster für diese Buchreihe gelten. Ge.

KRUEGER, F. R.
Physik und Evolution
Physikalische Ansätze zu einer Einheit der Naturwissenschaften auf evolutiver Grundlage
1984, 211 S., DM 46,–
Paul Parey Verlag, Hamburg und Berlin

Dies ist ein weiterer Band in der Buchreihe „Biologie und Erkenntnis". Der Autor macht dabei physikalische Ansätze zu einer Einheit der Naturwissenschaften auf evolutiver Basis und begründet, inwiefern die Physik als Grundlage der Biologie anzusehen ist. Dabei werden die wesentlichsten physikalischen Prinzipien der Evolution auch für den Nichtphysiker verständlich dargestellt. Mögliche Berührungspunkte mit Geisteswissenschaften werden andiskutiert. Evolution erscheint als kategorischer Imperativ der praktischen Vernunft. Ge.

SCHUBÖ, W. und UEHLINGER, H.-M.
SPSS^X Handbuch der Programmversion 2
1984, 493 S., DM 44,–
G. Fischer Verlag, Stuttgart–New York

Die Verbreitung von SPSS ist insbesondere darauf zurückzuführen, daß gute deutsche Handbücher vorlagen. So ist es nicht mehr als konsequent, daß nun auch für die neue Version SPSS^X ein neues Handbuch erstellt wurde, das dem geänderten Umfang und den weiteren Möglichkeiten bei den Datenstrukturen Rechnung trägt.

Auch dieses Handbuch wird sicher wieder ein nützliches Hilfsmittel bei der Verwendung des SPSS^X Programmpakets zur Datenanalyse für die große Zahl von Anwendern sein. Ge.

SPIESS, E. W. und RHEINGANS, F. G.
Einführung in das Programmieren von FORTRAN
7. Auflage
1985, 250 S., DM 28,–
Walter de Gruyter, Berlin – New York

Die Neuauflage dieser bewährten Einführung ist weitgehend eine Neubearbeitung und stützt sich auf FORTRAN 77, wobei speziell für Anfänger mit einem Subset (Teil-FORTRAN) gearbeitet, aber insgesamt das Gesamt-FORTRAN dargestellt wird.

Die Darstellungsart ist auch in dieser Auflage überzeugend und vermittelt am Rande Informationen über den Aufbau von Computern und u. a. auch über Probleme der Rechengenauigkeit. Da heute auch für viele PC's FORTRAN-Compiler verfügbar sind, darf man auch dieser Auflage wieder eine weite Verbreitung wünschen. Ge.